SPATIAL2

Spatial Data Methods for Environmental and Ecological Processes - 2nd Edition

PROCEEDINGS

EDITOR: Barbara Cafarelli

ENVIRONMETRICS

Contents

- **The deep-water rose shrimp in the Ionian Sea: a spatio-temporal analysis of zero-inflated abundance data, D'Onghia G., Maiorano P., Carlucci R., Tursi A, Pollice A., Ribecco N., Calculli C., Arcuti S.**

Climatology and Meteorology

- **A few links between the notion of Entropy and Extreme Value Theory in the context of analyzing climate extremes, Naveau P., Rietsch T., Guillou A., Merleau J.**
- **Geoadditive modeling for extreme rainfall data, Bocci C., Petrucci A., Caporali E.**
- **Spatio-temporal rainfall trends in southwest Western Australia, Liang K., Chandler R., Marra G.**
- **Stochastic Downscaling of Precipitation with Conditional Mixture Models, Carreau J., Vrac M.**

Disease Mapping and Environmental Exposure

- **A Bayesian Spatio-Temporal framework to improve exposure measurements combining observed and numerical model output, Pirani M., Gulliver J., Blangiardo M.**
- **A spatio-temporal model for cancer incidence data with zero-inflation, Musio M., Sauleau E.A.**
- **Generalized Estimating Equations for Zero-Inflated Spatial Count Data, Monod A.**
- **Poisson M-Quantile Geographically Weighted Regression on Disease mapping, Chambers R., Dreassi E, Salvati N**

Environmental Data Analysis

- **A software for optimal information based downsizing/upsizing of existing monitoring networks, Barca E., Passarella G., Vurro M., Morea A.**
- **Comparing SaTScan and Seg-DBSCAN methods in spatial phenomena, Montrone S., Perchinunno P., L'Abbate S., Ligorio C.**
- **Fire, earthquake, landslide, volcano, flood: first approach to a natural hazard map of Italy, Camporese R., Iandelli N.**
- **Spatio-Temporal Analysis of Wildfire Patterns in Galicia (NW Spain), Fuentes-Santos I., Gonzalez-Manteiga W., Marey-Pérez. M. F.**

GIS and Soil Sciences

- **Imputation strategy in spatial data, Martino L., Palmieri A.**
- **Multivariate geostatistical model to map soil properties at a region scale from airborne hyperspectral imagery and scattered soil field surveys: dealing with large dimensions, Monestiez P., Walker E., Gomez C., Lagacherie P.**
- **Optimal location and size for a biomass plant:application of a GIS methodology to the "Capitanata" district, Monteleone M., Cammerino A.R.B., lo Storto M.C.**
- **Population Density in a City, Abbate C., Salvucci G.**

Landscape Ecology and Natural Resource Management

- **Comparison of spatial statistics for identifying underlying process in forest ecology, Brown C., Illian J., Burslem D., Law R.**
- **Connectivity in a real fragmented landscape: distance vs movement model based approaches, Mairota P., Leronni V., Cafarelli B., Baveco J.M.**
- **Methodological study on pesticides in Alsatian groundwater, Musci F., Giasi C.I., de Fouquet C.**
- **The GIS approach to detect the influence of the fresh water inflows on the marine-coastal waters: the case of the Apulia Region (Italy) through standard monitoring**

data, Porfido A., Barbone E., La Ghezza V., Costantino G., Perrino V., Ungaro N., Blonda M.

Methods and Environmental Modelling

- Applying a new procedure for fitting a multivariate space-time linear coregionalization model, De Iaco S., Palma M., Posa D.
- Decision making for root disease control: a problem in reducing the nugget variance, Correll R.
- EM estimation of the Dynamic Coregionalization Model with varying coefficients, Finazzi F., Fassò A.
- Likelihood Inference in Multivariate Model-Based Geostatistics, Ferrari C., Minozzo M.

Proximal and Remote Sensing in Precision Agriculture

- A system for on-line measurement of key soil properties, Mouazen A.M., Kuang B., Quraishi M.Z.
- Modified Hot-Spot analysis for spatio-temporal data: a case study of the leaf-roll virus expansion in vineyards, Cohen Y., Sharon R., Sokolsky T., Zahavi T.
- Multimodal remote sensing for enhancing detection of spatial variability in agricultural fields, Alchanatis V., Cohen A., Cohen Y., Levi O., Naor A.
- The use of the geoadditive model with interaction in a Precision Agriculture context: a comparison of different spatial correlation structures, Cafarelli B., Crocetta C., Castrignanò A

Sampling Designs for Natural Studies

- On the design-based properties of spatial interpolation, Bruno F., Cocchi D., Vagheggini A.
- Relations between spatial design criteria, Mueller W.G., Waldl H.
- Simulation-based optimal design for estimating weed density in agricultural fields, Bel L., Parent E., Makowski D.
- The dramatic effect of preferential sampling of spatial data on variance estimates, Clifford D., Kuhnert P., Dobbie M, Baldock J., McKenzie N., Harch B., Wheeler I., McBratney A.

Space-Time Surveillance for Public Health

- Modeling malaria incidence in Sucre state, Venezuela using a Bayesian approach, Villalta D., Guenni L., Rubio Y.
- Prediction of cancer mortality risks in spatio-temporal disease mapping, Goicoa T., Ugarte M.D., Militino A.F., Etxeberria J.
- Predictive assessment of a non-linear random effects model for space-time surveillance data, Paul M., Held L.
- Selective Inference in Disease Mapping, Catelan D., Biggeri A.

Space-Time Surveillance of Natural Assets

- A seismic swarm as a dynamic ergodic stochastic process: a case study of the L'Aquila's earthquake in 2009, Coli M.
- Geostatistical modeling of ice content within the "Glacier Bonnard" (Switzerland), Jeannee N., Faucheux C., Bardou E., Ornstein P.
- Is space-time interaction real or apparent in seismic activity?, Rotondi R., Varini E.
- Spatio-temporal modelling for avalanche risk assessment in the North of Italy, Nicolis O., Assuncao R.

5. Poster session

Agriculture, biodiversity, groundwater pollution and hydrogeology

- **A data driven model for spatio-temporal estimation of shallow water table depth in soils, Ungaro F., Calzolari C.**
- **Assessment and modelling of spatial variability of the soil factors potentially affecting groundwater nitrate contamination in two agricultural areas of Molise Region (Southern Italy), Colombo C., Palumbo G., Sollitto D., Castrignanò A.**
- **Assessment of Spatial and Temporal Within-Field Soil Variability by Using Geostatistical Techniques, Castrignanò A., Cucci G., Diacono M., De Benedetto D., Lacolla G., Troccoli A.**
- **CYCAS-MED project: analysis at regional and local scale of climate change impacts on cereals yield in Morocco, Bodini A., Entrade E., Cesaraccio C., Duce P., Zara P., Dubrovsky M.**
- **Geostatistical analysis and mapping of hydrocarbon pollutants in soils, de Fouquet Chantal**
- **Geostatistical analysis of groundwater nitrates distribution in the Plaine d'Alsace, Spacagna R.L., De Fouquet C., Russo G.**
- **Influence of different olive grove management on spider diversity, Loverre P., Addante R, Calculli C.**
- **Landcover classification of agricultural sites using multi-temporal COSMO-Skymed data, Satalino G., Balenzano A., Belmonte A., Mattia F., Impedovo D.**
- **Multidimensional analysis of data from Bari Harbour: a GIS based tool for the characterization and management of bottom sediments, Dellino P., Mele D., Mega M., Pagnotta E., De Giosa F., Taccardi G., Ungaro N., Costantino G.**
- **Multivariate statistical analyses for the source apportionment of groundwater pollutants in Apulian agricultural sites, Ielpo P., Cassano D., Lopez A., Abbruzzese De Napoli P., Pappagallo G, Uricchio V.F.**
- **Structural changes in seismic activity before large earthquakes, Gallucci M., Petrucci A.**
- **Using environmental metrics to describe the spatial and temporal evolution of landscape structure and soil hydrology and fertility, Pascual Aguilar J. A., Sanz Garcia J., de Bustamante Gutierrez I., Kallache M.**

Air quality and disease mapping

- **A comparison between hierarchical spatio-temporal models in presence of spatial homogeneous groups: the case of Ozone in the Emilia-Romagna Region, Bruno F., Paci L.**
- **A multilevel multimember model for smoothing a disease map of lung cancer rates, Bartolomeo N., Trerotoli P., Serio G.**
- **A spatio-temporal model for air quality mapping using uncertain covariates, Cameletti M., Ghigo S., Ignaccolo R.**
- **African dust contribution on the PM10 daily exceedances occurred in Apulia region, Angiuli L., Giua R., Loguercio Polosa S., Morabito A.**
- **Health impact assessment of pollution from incinerator in Modugno (Bari), Galise I, Serinelli M., Bisceglia L., Assennato G.**
- **Local scoring rules for spatial processes, Dawid P., Musio M.**
- **Measuring Urban Quality of Life Using Multivariate Geostatistical Models, Michelangeli A., Ferrari C., Minozzo M.**
- **Multivariate and Spatial Extremes for the Analysis of Air Quality Data, Padoan S., Fassò A.**

- **Pulmonary Tuberculosis and HIV/AIDS in Portugal: joint spatio-temporal clustering under an epidemiological perspective, Nunes C., Briz T., Gomes D., Filipe P.A.**
- **Spatial diffusion and temporal evolution of PCDD/Fs, PCBs and PAHs congener concentrations in the ambient air of Taranto: an analysis based on the duality diagram approach, Pollice A., Esposito V.**
- **Spatial disaggregation of pollutant concentration data, Horabik J., Nahorski Z.**
- **Spatial representativeness of an air quality monitoring station. Application to NO2 in urban area, Beauchamp M., Malherbe L., Létinois L., de Fouquet C.**
- **Statistical investigations on PAH concentrations at industrial sampling site, Amodio M., Andriani E., Dambruso P.R., de Gennaro G., Demarinis Loiotile A., Di Gilio A., Trizio L., Assennato G., Colucci C., Esposito V., Giua R., Menegotto M, Spartera M.**
- **Tapering spatio temporal models, Fassò A., Finazzi F., Bevilacqua M.**

Climatology and meteorology and sampling design

- **A Methodology for Evaluating the Temporal Stability of Spatial Patterns of Vineyard Variation, Gambella F., Dau R., Paschino F., Castrignanò A., De Benedetto D.**
- **Alternative approaches for probabilistic precipitation forecasting, Bruno F., Cocchi D., Rigazio A.**
- **Comparison of Calibration Techniques for Limited-Area Ensemble Precipitation Forecast Using Reforecasts, Diomede T., Marsigli C., Montani A., Paccagnella T.**
- **Functional boxplots for summarizing and detecting changes in environmental data coming from sensors, Romano E., Balzanella A., Rivoli L.**
- **Information, advice, friendship, notes and trust network: evidence on learning from classmate, Zavarrone Emma, Vitali Agnese**
- **Optimal spatial design for air quality measurement surveys: what criteria?, Romary T., de Fouquet C., Malherbe L.**
- **Point-process statistical analysis for the ECMWF Ensemble Prediction System, Nerozzi F.**

Ecology, conservation and natural resources management

- **Combining geostatistics and process-based water quality model to improve estimation along a stream network. Example on a stretch of the Seine River, de Fouquet C., Polus-Lefèbvre E., Flipo N., Poulin M.**
- **Landscape impacts of photovoltaic plants on the ground: a case-study through the application of rendering techniques, Robles N., Primerano R., Perrino V., Blonda M.**
- **Marine spatial planning in Apulia (Italy): Reconciling seagrass conservation with the multiple use of coastal areas, Fraschetti S., Lembo G., Tursi A., D'Ambrosio P., Terlizzi A., De Leo F., Paes S., Guarnieri G., Bevilacqua S., Boero F.**
- **Regional estimation method of rivers low flow from river basin characteristics, Rossi G., Caporali E.**
- **Spatial Analysis of some soil physicochemical properties in mountainous massif of Sico, Portugal, Torres O. M., Neves M. M., Gomes D. P.**
- **Spatial and auto correlation of ecological change: disturbance and perturbation analysis in Circeo National Park (south Latium, Italy), Galante G., Cotroneo R., Mandrone S., Strafella I.**
- **Spatial diversity in a "zoom-lens": Analysing ecological communities through weighted spatial scales, Studeny A. C., Brown C., Illian J.B.**

- **Spatio-temporal changes of biodiversity indices in the bathyal demersal assemblages of the Ionian Sea, Maiorano P., Giove A., Minerva M., Sion L., D'Onghia G , Pollice A., Ribecco N., Muschitiello C.**
- **Spatio-temporal variability in stream flow status: Candelaro river case study, De Girolamo A.M., Calabrese A., Pappagallo G., Santese G., Lo Porto A., Gallart F, Prat N., Froebrich J.**
- **Statistical assessment of the plant protection level within protected areas (PA) based on remote sensing products, Menconi M.E., Pacicco C.L.**
- **Statistical calibration of the Carlit index in the Pontine Island of Zannone, Jona Lasinio G., Tullio M.A., Abdelahad N., Scepi E., Sirago S., Pollice A.**
- **Statistical issues in the assessment of urban sprawl indices, Cocchi D., Altieri L., Scott M., Ventrucci M., Pezzi G.**
- **Using spatial statistics tools on remote-sensing data to identify fire regime linked with savanna vegetation degradation, Jacquin A., Chéret V., Goulard M., Sheeren D.**

Environmental risk assessment

- **A methodology for assessing the spatial distribution of static wildfire risk over wide areas: the case studies of Liguria and Sardinia (Italy), Bodini A., Entrade E., Cossu Q. A., Canu S., Fiorucci P., Gaetani F., Paroli U.**
- **A new procedure for fitting a multivariate space-time linear coregionalization model, De Iaco S., Palma M., Posa D.**
- **Bayesian hierarchical models: An analysis of Portugal road accident data, Ribeiro C., Turkman A. A., Cardoso J.L.**
- **Electrical Resistivity Measurements for Spatial Soil Moisture Variability Estimation, Calamita G., Luongo R., Perrone A., Lapenna V., Piscitelli S., Straface S.**
- **Geostatistics and GIS: tools for environmental risk assessment, Maggio S., Cappello C., Pellegrino**
- **How to estimate anisotropic attenuation exploiting prior isotropic knowledge, Rotondi R., Zonno G.**
- **Natural radioactivity distribution and soil properties: a case study in southern Italy, Guagliardi I., Ricca N., Cipriani M.G., Civitelli D., Froio R., Gabriele A.L., Buttafuoco G., De Rosa R.**
- **Screening level risk assessment for phenols in surface water of three rivers in Tianjin, China, Zhong W., Wang D., Wang Z., Zhu L.**
- **Spatial Dynamic Factor Models with environmental applications, Valentini P., Ippoliti L. , Gamerman D.**
- **Spatial Point Processes Applied to the Study of Forest Fires in Portugal, Pereira P.S., Turkman K.F.**
- **Spatio-Temporal Analysis of Forest Fires in Portugal, Dias M. I., da Silva G.L.**

6. Scientific Program

7. Spatial Café Program

8. Spatial2 – Poster

9. Cover A - B

Preface

This book collects the proceedings of the International Conference "Spatial Data Methods for Environmental and Ecological Processes - 2nd Edition", the 2011 European Regional Conference of The International Environmetrics Society, satellite of the 58th World Statistics Congress of the International Statistical Institute (ISI).

The main scope of the conference is exchanging past results and new ideas among researchers with different scientific backgrounds, all working on spatial and spatio-temporal environmental problems.

The conference is structured into five plenary sessions, twelve specialized sessions and a poster session, as follows:

Plenary sessions:

- Climatology and Meteorology
- Ecology and Water Analysis
- Ensemble Forecasts
- Sampling and Accurate Predictions for Environmental Management
- Spatial Functional Data

Specialized sessions:

- Air Quality
- Animal and Plant Ecology
- Climatology and Meteorology
- Disease mapping and Environmental Exposure
- Environmental Data Analysis
- GIS and Soil Sciences
- Landscape Ecology and Natural Resource Management
- Methods and Environmental Modelling
- Proximal and Remote Sensing in Precision Agriculture
- Sampling Designs for Natural Studies
- Space-time Surveillance for Public Health
- Space-time Surveillance of Natural Assets

Main themes of the poster session

- Agriculture, Biodiversity, Groundwater Pollution and Hydrogeology
- Air Quality and Disease Mapping
- Climatology and Meteorology and Sampling design
- Ecology, Conservation and Natural Resources Management
- Environmental Risk Assessment

The poster discussion was held during a "Spatial Café"

The Spatial Café was organized in five discussion tables. For each table two facilitators were chosen to stimulate and organize the posters discussion.

The Conference's Scientific Committee tailored the program to provide fruitful interactions among various research fields, under the common heading of “spatial analysis”. This was very clear during the course of the conference, as communication among participants both from Italy and abroad, from universities and research centers, and most importantly, among statisticians and researchers from other subject areas, was facilitated by a charming, very friendly atmosphere.

This Volume of Proceedings contains 110 short papers and abstracts that were presented during the conference and is articulated in three parts, each corresponding to a session held in the conference. All published papers were submitted to a refereeing process. The refereeing process has been attended by the Scientific and Organizing Committees.

The Scientific and Organizing Committees are very grateful to the University of Foggia, the University of Bari, the Fondazione Cassa di Risparmio di Puglia, The International Environmetrics Society, the International Statistical Institute, the Società Italiana di Statistica, the CRA-CSA of Bari, the Agenzia Regionale per la Prevenzione e la Protezione dell’Ambiente - Puglia - and the GRASPA research group for supporting the organization of the conference and allowing us to publish this volume.

In quality of Scientific Committee and Organizing Committee Presidents, we would like to thank the members of the Scientific Committee (Liliane Bel, Annamaria Castrignanò, Corrado Crocetta, Alessandro Fassò, Giovanna Jona Lasinio, Alessio Pollice and Marian Scott) and of the Organizing Committee (Barbara Angelillis, Francesca Bruno, Rosalba Ignaccolo, Giovanna Jona Lasinio, Alessio Pollice and Alessia Spada) for their outstanding work and all the participants to the conference for their contributions.

Daniela Cocchi, President of the Scientific Committee
Barbara Cafarelli, President of the Organizing Committee

Global temperature analysis with non-stationary random field models

Finn Lindgren, Håvard Rue
Norwegian University of Science and Technology, finn.lindgren@math.ntnu.no

Peter Guttorp
University of Washington and Norwegian Computing Center

Abstract: Analysis of regional and global mean temperatures based on instrumental observations has typically been based on aggregating temperature measurements to grid cells. Due to the uneven data coverage, this makes analysis of the associated uncertainties difficult. We here present an alternative model based approach, where the climate and weather are modelled as random fields generated by a stochastic partial differential equation. Using the efficient Markov representations developed by Lindgren et al. (2011), direct numerical optimisation and integration with the R-INLA software provides Bayesian temperature reconstructions and associated uncertainties.

Keywords: Global temperature analysis, Stochastic partial differential equation, Gaussian Markov random field

1 Introduction

When analysing past observed weather and climate, the Global Historical Climatology Network (GHCN) data set (Peterson and Vose, 1997) is commonly used. The data spans the period 1702 through 2010, though counting, for each year, only stations with no missing values, yearly averages can be calculated only as far back as 1835. The GHCN data is used to analyse regional and global temperatures in the GISS (Hansen et al., 1999) and HadCRUT3 (Brohan et al., 2006) global temperature series, together with additional data such as ocean based sea surface temperature measurements. Differing in detail, the analyses aggregate the data into grid boxes, which are combined into global averages. To reduce the influence of station specific effects, the methods are based on the temperature anomalies, defined as the difference in weather to the local climate, the latter defined as the average weather over a 30 year reference period. Due to the difficulty of assessing the statistical uncertainty of the resulting estimates, we instead choose to construct a stochastic model for the climate and anomalies, based on a non-stationary stochastic partial differential equation.

2 Model

I order to avoiding the computational difficulties associated with calculations based on covariance matrices, we use the link between the stochastic partial differential equation (SPDE) formulation of Matérn fields and Gaussian Markov random fields (GMRFs), as developed by Lindgren et al. (2011). Together with the INLA method (Rue et al., 2009) this allows us to perform a fully Bayesian analysis in a fraction of the time required by a traditional MCMC approach.

The climate (or expected weather) is $\boldsymbol{\mu}$, the yearly anomalies are $\boldsymbol{x}_t$, and the observations are $\boldsymbol{y}_t$. The anomalies are taken as solutions to the SPDE

$$\left(\kappa^2(\boldsymbol{u}) - \Delta\right)\left(\tau(\boldsymbol{u})x_t(\boldsymbol{u})\right) = \mathcal{W}(\boldsymbol{u}), \quad \boldsymbol{u} \in \mathbb{S}^2, \tag{1}$$

where $\mathcal{W}$ is a white noise process, Δ is the Laplacian, and κ and τ are spatially varying parameters. The prior distribution for the climate field is chosen as approximate solutions to the SPDE $\Delta\mu(\boldsymbol{u}) = \sigma_\mu\mathcal{W}(\boldsymbol{u})$, which are intrinsic random fields. The model is governed by a parameter vector $\boldsymbol{\theta} = \{\boldsymbol{\theta}_\kappa, \boldsymbol{\theta}_\tau, \boldsymbol{\theta}_s, \theta_\epsilon\}$, where $\boldsymbol{\theta}_\kappa$ and $\boldsymbol{\theta}_\tau$ controls the non-stationary dependence structure of the anomalies.

Introducing *observation matrices* $\boldsymbol{A}_t$, that extract the nodes from $\boldsymbol{x}_t$ for each observation, the full model is given by

$$(\boldsymbol{\mu}|\boldsymbol{\theta}) \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{Q}_{\boldsymbol{\mu}}^{-1}), \tag{2}$$

$$(\boldsymbol{x}_t|\boldsymbol{\theta}) \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{Q}_{\boldsymbol{x}}^{-1}), \tag{3}$$

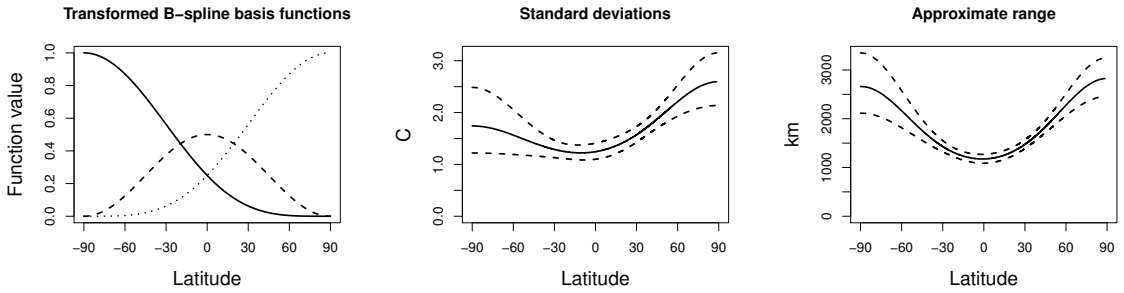
$$(\boldsymbol{y}_t|\boldsymbol{\mu}, \boldsymbol{x}_t, \boldsymbol{\theta}) \sim \mathrm{N}(\boldsymbol{A}_t(\boldsymbol{\mu} + \boldsymbol{x}_t) + \boldsymbol{S}_t\boldsymbol{\theta}_s, \boldsymbol{Q}_{\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{x}}^{-1}), \tag{4}$$

where $\boldsymbol{S}_t\boldsymbol{\theta}_s$ are station specific effects (elevation), and the $\boldsymbol{Q}_\cdot$ matrices are the precision matrices corresponding to each conditional distribution, obtained with the finite element method (Lindgren et al., 2011).

3 Results

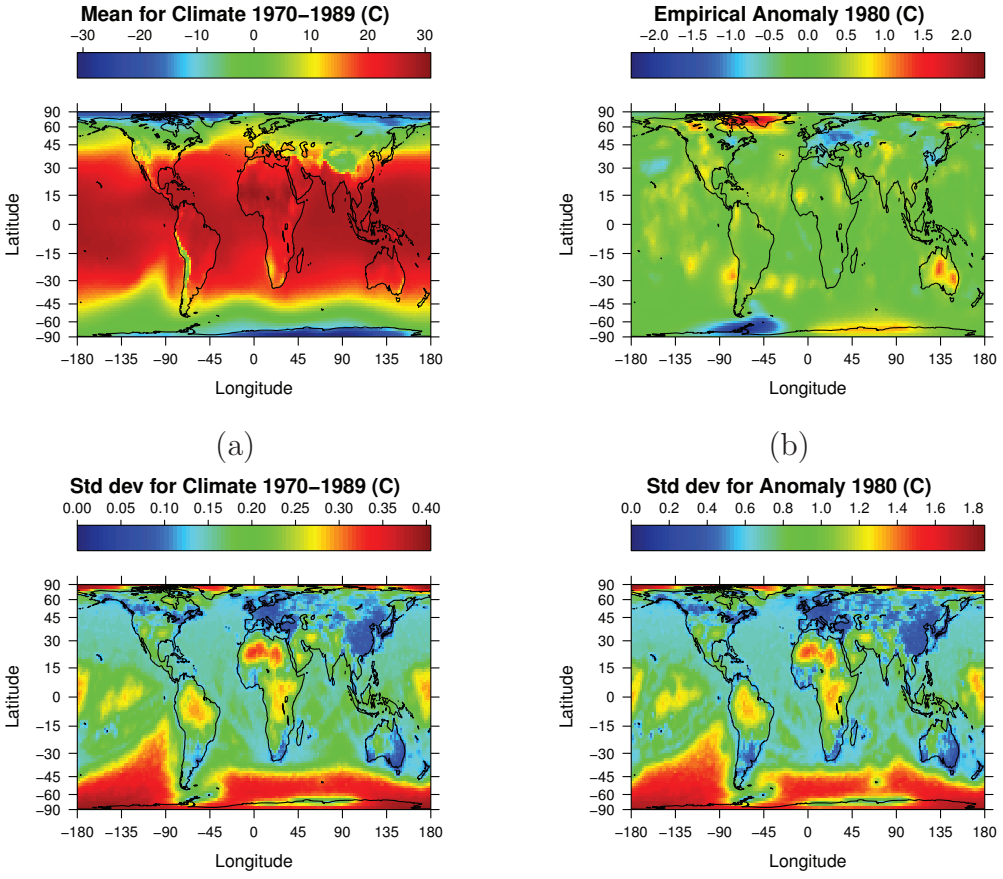
We implemented the model using `R-INLA`. The Bayesian analysis draws all its conclusions from the properties of the posterior distributions of $(\boldsymbol{\theta}|\boldsymbol{y})$, $(\boldsymbol{\mu}|\boldsymbol{y})$, and $(\boldsymbol{x}|\boldsymbol{y})$, so that all uncertainty about the weather anomaly $\boldsymbol{x}_t$ is included in the distribution for the model parameters $\boldsymbol{\theta}$, et cetera. Since $(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ is Gaussian, the Bayesian integration results are only approximate with regards to the numerical integration of the covariance parameters $(\boldsymbol{\theta}_\kappa, \boldsymbol{\theta}_\tau, \boldsymbol{\theta}_\epsilon)$. Due to the large size of the data set, the initial analysis is based on data only from the period 1970 through 1989, and the analysis took approximately one hour on a 12 core Linux system.

The spatial covariance parameters are harder to interpret individually, but we instead show the resulting spatially varying field standard deviations and correlation ranges in Figure 1, including pointwise 95% credible intervals. Both curves show a clear dependence on latitude, with both larger variance and correlation range near the poles, compared with the equator.



(a) (b) (c)

Figure 1: Three transformed B-spline basis functions of order 2 (a), and approximate 95% credible intervals for (b) standard deviation and (c) correlation range of the yearly weather, as functions of latitude.



(a) (b) (c) (d)

Figure 2: Posterior means for the empirical 1970–1989 climate (a) and for the empirical mean anomaly 1980 (b), together with the corresponding posterior standard deviations in (c) and (d). The climate includes the estimated effect of elevation. An area-preserving cylindrical projection is used.

In Figure 2(a) and (b), the posterior expectation of the empirical climate, $\mathrm{E}(\boldsymbol{\mu}|\boldsymbol{y})$, is shown (with the estimated effect of elevation added), together with the posterior expectation of the temperature anomaly for 1980, $\mathrm{E}(\boldsymbol{x}_{1980}|\boldsymbol{y})$. The spatial dependence model was based on the GHCN data, but these Kriging estimates also include ocean-based data. A preliminary analysis indicates that the dependence structure is different for land and ocean, which can be handled by adding appropriate basis functions to the κ and τ models. The pre-gridded ocean data is also a good example of how the observation matrix $\boldsymbol{A}_t$ can solve the problem of "misaligned" data, since it decouples the spatial model from the data locations, allowing arbitrary linear measurement equations from one spatial model.

References

Brohan, P., Kennedy, J., Harris, I., Tett, S., and Jones, P. (2006). Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. *Journal of Geophysical Research*, 111.

Hansen, J., Ruedy, R., Glascoe, J., and Sato, M. (1999). GISS analysis of surface temperature change. *Journal of Geophysical Research*, 104:30997–31022.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach approach (with discussion). *Journal of the Royal Statistical Society, Series B*. In press.

Peterson, T. and Vose, R. (1997). An overview of the Global Historical Climatology Network temperature database. *Bulletin of the American Meteorological Society*, 78(12):2837–2849.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71(2):319–392.

Methods for climate change detection and attribution

Aurélien Ribes
CNRM - GAME, Météo France - CNRS, aurelien.ribes@meteo.fr

Abstract: Detection and attribution (D&A) have played a central role within the assessment of the human influence on climate and within IPCC's reports. Detection involves the statistical demonstration that a change has happened within climatic observations. Attribution consists in assessing the respective contributions of one or several causes to some observed change. Both require the use of climate model simulations, and are based on spatial or spatio temporal patterns of change. This paper provides a very short presentation of the classical "optimal fingerprint" method for D&A. Some recent developments, regarding the use of "error in variable" are introduced. Some of the challenging aspects of the method will be discussed too, in particular regarding the very large dimension of the typical datasets used.

Keywords: Climate change, detection, attribution, linear model, high dimension.

1 Introduction

Detection and attribution (D&A) have played a central role within the assessment of the human influence on climate and within IPCC's reports. Detection involves the statistical demonstration that a change has happened within climatic observations. Attribution consists in assessing the respective contributions of one or several causes to some observed change. Both are based on the characterisation of the spatial or spatio-temporal pattern of change corresponding to each physically plausible cause. However, specific tools from spatial statistics have been poorly used on that theme.

This paper aims primarily at giving a state of the art picture of some of the concepts, statistical tools, and current challenges in D&A analysis. The secondary attempt is to shortly discuss both difficulties and potential benefits of using spatial statistics tools.

Introduction of D&A first requires to introduce some concepts used in climate sciences. Climatologists use to first define their subject of study: the climate system. It includes the atmosphere, the ocean, and several other components (see IPCC, 2007). This system is influenced by several boundary conditions (e.g. the solar activity, the chemical composition of the atmosphere), usually referred to as *external forcings*, that may impact its state or dynamics. However, the variables used for describing the state of the system show some variability, even under fixed boundary

conditions. This variability is called *internal variability*, and corresponds to the kind of variability expected while the climate is not changing.

Statistical D&A requires to have some knowledge on two parameters: first, the statistical properties or the distribution of the internal variability, and second, the expected response of the climate system to a given external forcing. Physically-based climate models are usually used for evaluating both objects instead of e.g. parametric models. Indeed, internal variability involves very specific spatial patterns and a large set of spatial scales that may hardly be accounted for in a parametric model. Instead, the use of climate model allows the evaluation from our physical understanding. D&A then requires careful comparison between observed changes and outputs from climate models.

2 Optimal fingerprint method

The more classical approach for climate change D&A is usually referred to as the *optimal fingerprint* method. This method has been gradually introduced at the end of the 90's (Hasselmann, 97, Hegerl et al., 97, Allen & Tett, 99). The latter presents this method as a linear regression of the observed climate time-series on the expected responses to the external forcings :

$$Y = \sum_{i=1}^{I} \beta_i g_i + \varepsilon, \tag{1}$$

where Y are the observations, β_i are unknown scaling factors, g_i is the expected response of the system to the i-th external forcing (as simulated by one or several climate model), and ε denotes the internal variability. In Eq. (1), Y is usually a spatio-temporal vector, Y_i typically consisting of the average of the temperature over a region, and a decade. g_i and ε have the same dimension and structure as Y.

Model (1) basically assumes that climate models have some accuracy at simulating the spatio-temporal pattern of the response to each external forcing, whereas they may fail at simulated the proper amplitude of that response. Within model (1), detection of a change associated to the forcing i corresponds to the rejection of the null hypothesis "$\beta_i = 0$". Attribution, in addition to the detection, requires to show that the observed response is consistent with the expected one, or equivalently, that the null hypothesis "$\beta_i = 1$" cannot be rejected.

Assuming that $C = \mathrm{Cov}(\varepsilon)$ is known, for example from climate models simulations, the computation of maximum likelihood estimate (MLE) for β is easy:

$$\widehat{\beta} = (G'C^{-1}G)^{-1}G'C^{-1}Y, \tag{2}$$

where $G = [g_1, \ldots, g_I]$. Under the same assumption, the distribution of the MLE is known, so as hypothesis testing on β is easy to perform.

Some refinement of the method has been introduced by Allen and Stott (2003) and Huntingford (2006), in order to take into account the uncertainty at simulating

the spatio temporal patterns g_i. In that case, the uncertainty may come from internal variability (within the climate model simulation), or multi-model uncertainty. The main statistical model is then slightly changed to

$$Y = \sum_{i=1}^{I} \beta_i (g_i + \nu_i) + \varepsilon, \tag{3}$$

where ν_i represents the uncertainty on g_i. Assuming, similarly to ε, that $\Sigma = \mathrm{Cov}(\nu)$ is known, the optimal estimate of β may be derived by using a *Total Least Square* (TLS) procedure instead of the *Ordinary Least Square* technique involved in the MLE mentioned before.

Such methods have led to one of the important figures of the last IPCC report that deals with the quantification of the contribution of several external forcing to the observed warming (Figure 9.6, IPCC, 2007).

3 Estimation of C and high-dimension

The method presented before assumes that C (and Σ) is known, while, in a real-life problem, it is not. Several difficulties arise from the estimation of C, that is usually done from a *control* runs (i.e. climate simulations without any change in the external forcings). We here will focus in the problem related to the high-dimension of the typical global temperature datasets.

Current datasets are providing homogenised temperatures on a 5° x 5° grid, that results in 2592 grid-points in space. D&A study typically consider a 50-yr period in time, decomposed in 5 decades. The dimension of Y is then close to 13000. Note that missing values will likely decrease this number, but won't change the typical size of, say 10^4. Consequently, C is a 10^4 x 10^4 matrix, that has to be estimated from available control runs, that are typically covering 10^4 years (when considering together control runs from various models). Classical covariance matrix estimates being very poor in such cases, the dimensionality needs to be reduced.

Two approaches have been mainly used in order to reduce this dimension while focusing on the large spatial scales. First, global temperatures have been projected onto some first spherical harmonics (e.g. Stott, 2006). Second, particularly at the regional scale (where the dimension of the dataset is smaller but remains too high), data have been projected onto the first principal components (e.g. Zwiers, 2003). In both cases, projection may reduce the accuracy of the β estimates (there are no results of optimality), and requires to choose the reduced dimension (i.e. the number of spherical harmonics or principal components), what may be sensitive.

One possible alternative consists in using a regularised estimate of the covariance matrix C, that is a linear combination of the empirical covariance matrix estimate $\widehat{C}$ and the identity (Ribes et al., 2009) :

$$\widetilde{C} = \gamma \widehat{C} + \rho I. \tag{4}$$

Such an estimate has been shown to be more accurate than $\widehat{C}$ in high-dimension (Ledoit and Wolf, 2004). To plug (4) into (2) also leads to an improved estimate of β in the context of high dimension data.

This approach may help the estimation of β, but no result of optimality has been proved. As a consequence, the problem of efficiently estimating β in the context of high dimension dataset is still open. One potentially attractive way may be to use the spatio-temporal structure of Y in order to improve the estimation of C.

4 Concluding remarks

D&A deal with one key-question regarding climate change, that is the quantification of the human contribution to the current warming. While initially based on a simple linear model, D&A involve some recent statistical tools and also provide some challenging questions, in particular related to the high dimension of the corresponding datasets.

References

Allen M. R. and Tett S. F. B. (1999) Checking for model consistency in optimal fingerprinting, *Climate Dynamics*, 15, 419-434.

Allen M. R. and Stott P. A. (2003) Estimating signal amplitudes in optimal fingerprinting, Part I: Theory, *Climate Dynamics*, 21, 477-491.

Hasselmann K. (1997) Multi-pattern fingerprint method for detection and attribution, *Climate Dynamics*, 13, 601-612.

Hegerl G. C. et al. (1997) Multi-fingerprint detection and attribution of greenhouse-gas and aerosol-forced climate change, *Climate Dynamics*, 13, 613-634.

Huntingford C. et al. (2006) Incorporating model uncertainty into attribution of observed temperature change. *Geophysical Research Letters*, 33, L05710.

IPCC (2007) Climate change 2007: the physical basis. Contribution of working group 1 to the fourth assessment report of the international panel on climate change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 996pp.

Ledoit O. and Wolf M. (2004) A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis*, 88, 365-411.

Ribes A., Azaïs J.-M. ans Planton S. (2009) Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate, *Climate Dynamics*, 33, 707-722.

Stott P. A. et al. (2006) Observational constraints on past attributable warming and predictions of future global warming, *Journal of Climate*, 19, 3055-3069.

Zwiers F. W. and Zhang X. (2003) Towards regional scale climate change detection, *Journal of Climate*, 16, 793-797.

Assessing Temporal and Spatial Change in Nutrients for Large Hydrological Areas

Miller, C., Magdalina, A., Bowman, A.W., Scott, E.M. and Lee, D.
School of Mathematics and Statistics, University of Glasgow, UK
Claire.Miller@glasgow.ac.uk

Willows, R., Burgess, C., Pope, L and Johnson D.
Environment Agency, Evidence Directorate, UK

Abstract: Regulatory bodies, such as the Environment Agency of England & Wales, regularly monitor river surface water to assess quality. Maintaining and improving quality is important for society but is also a necessary requirement to comply with European directives. Spatiotemporal additive models for nutrients in hydrological areas in England & Wales are presented to assess and describe spatial and temporal trends over the past 20 to 40 years.

Keywords: spatiotemporal, smoothing, nutrients

1 Introduction

Previous modelling of nutrients within English & Welsh rivers has been carried out at individual monitoring locations in order to investigate trends over time and the effect of the contributing area at individual locations. However, for future nutrient policy decisions, there is a need to understand how historical patterns of water quality are described by catchment-scale influences rather than at individual sites. Spatiotemporal additive models have been developed for Large Hydrological Areas (LHAs) to investigate and describe nutrient trends on a catchment-wide basis.

2 Materials and Methods

2.1 The Data

There are 59 LHAs in England & Wales that contain independent river networks and their associated catchments. Monitoring locations within each LHA are associated with Water Framework Directive (WFD) waterbodies and each LHA consists of a number of such areal units. Orthophosphate (OP) and Total Oxidised Nitrogen (TON), mg/l, have been monitored on approximately a monthly basis by the Environment Agency of England & Wales over a period of 20 to 40 years at monitoring locations within each of the LHAs.

For this paper, the OP data in the Severn LHA, see Figure 1 (top left), will be investigated. These data span the time period 1971-2009 and have been aggregated within waterbodies. The OP data have been transformed using natural logs, to stabilise the variability throughout time, and measurements that were flagged as being below the limit of detection have been treated as censored observations and imputed (Helsel, D.R., 2005).

In order to investigate the effect of catchment covariates on nutrient levels, monitoring locations were selected which have no monitoring locations in further upstream waterbodies. This enables information from all local land area draining to the waterbody (contributing land) to be incorporated in the covariate value for an individual waterbody ensuring that contributed areas do not overlap and that there is little spatial correlation between measurements for a particular covariate.

The possible catchment covariates of interest are long-term (1961-1990) average base flow index (BFI) and discharge, total annual population, monthly total rainfall, fertiliser yearly application rates, land cover variables, livestock variables, crops, Agricultural Land Classification (ALC), slope of the land and soil type. All covariates have been aggregated within the 173 waterbodies of interest. Many of the continuous covariates have been log transformed to make their distributions more symmetric. Categorical variables were used for: slope (gentle to very steep), ALC (high quality agricultural to low quality grazing and non-agricultural land) and soil type (light to heavy), and for land cover, land use and population the hectares/counts have been standardised by the size of the contributing area.

2.2 Statistical Modelling

Model (1) was fitted to describe the relationships between the response of $\log_e$(OP) and all possible covariates: spatial and temporal trend and seasonality and the catchment covariates. A model which excludes the first three smooth terms of Model (1) was also fitted to investigate the relationships, and the percentage of variability explained, using only the catchment covariates, Model (2).

$$\begin{aligned} y \;=\; & \alpha + s(\text{Easting}, \text{Northing}) + s(\text{Year.month}) + s(\text{month}) + s(\text{discharge}) + s(\text{BFI}) \\ & + s(\text{land use}) + s(\text{land cover}) + \beta_{\text{ALC}_j} + \gamma_{\text{slope}_k} + \delta_{\text{soil}_l} + s(\text{rainfall}) \\ & + s(\text{population}) + s(\text{fertiliser}) + \epsilon \end{aligned} \tag{1}$$

where y is $\log_e$(OP), $s()$ is a smooth function, Year.month is decimal year and the errors (ϵ) are assumed to be $N(0, \sigma^2)$ and independent. For land use and land cover a series of different covariates are included individually such as potatoes, field vegetables, cows, etc. and the levels of the categorical variables are $j = 2, \ldots, 6$, $k = 2, \ldots, 4$, and $l = 2, \ldots, 8$. The degree of smoothing has been constrained to allow a maximum of 6 degrees of freedom for each univariate component to aid interpretation.

Functions from the `sm` library, see Bowman & Azzalini (1997) for details, and the `gam` function in the `mgcv` library, see Wood (2006) for full details, of `R` were used to fit these models. For the models that incorporate many covariates there is unlikely to be much spatiotemporal correlation remaining in the residuals and hence the assumption of independence appears appropriate. However, Moran's I Test in the `spdep` package of `R` and temporal variograms were used to check this assumption. If necessary, the methods of analysis can be modified to incorporate spatiotemporal correlation.

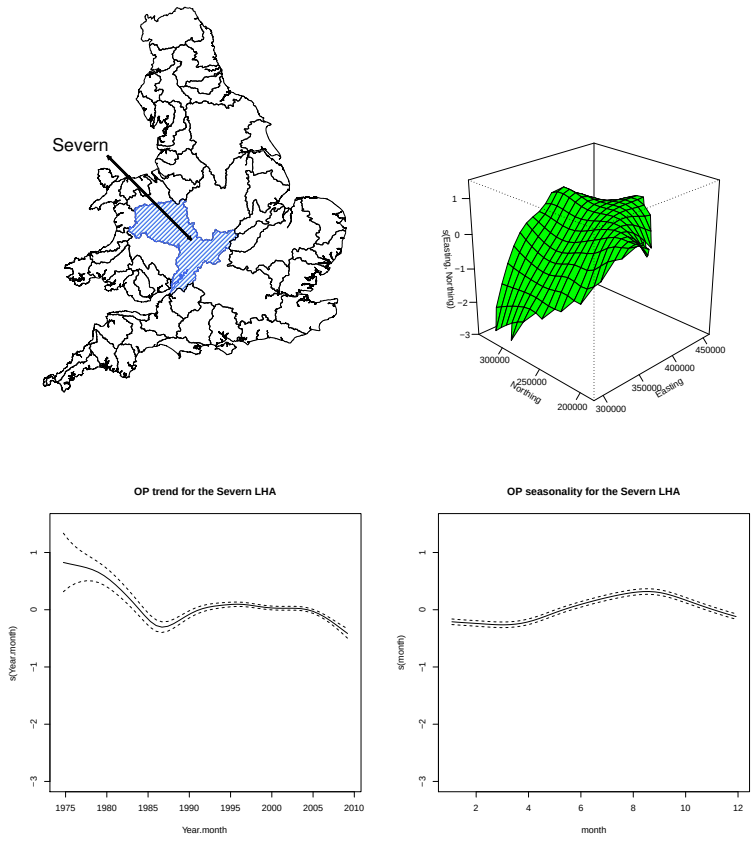


Figure 1: Map of Large Hydrological Areas in England & Wales with the Severn area highlighted (top left). For the Severn area: spatial trend (top right), temporal trend (bottom left), seasonal pattern (bottom right). The dashed lines indicate ± 2 standard errors.

3 Results

Figure 1 (top right and bottom panels) displays the spatial pattern, temporal trend and seasonality, respectively, for the Severn LHA. It highlights that the largest change for this area is spatial, followed by a smaller change over time and a small seasonal signal. Model (1) explains 70% of the variation in $\log_e(\text{OP})$, see Table 1,

with 66% of the variability explained using only the catchment covariates, Model (2). Therefore, the catchment covariates are usefully explaining the trends and seasonality in the area. Applying Moran's I and temporal variograms to the residuals from Model (1) suggests that there is very little evidence of spatial or temporal correlation remaining after incorporating all covariates.

In general, covariates are statistically significant as a result of the large amount of data. Many of the variables individually only explain a small proportion of the variability and hence it is difficult to reduce the number of covariates in the model. However, there are various possible combinations of a smaller subset of covariates that explain a reasonable amount of the variability. For example, reducing the number of catchment covariates from 23 to 8: ALC, soil, discharge, cattle, pigs, poultry, "other animals" and cumulative rainfall, still explains 46% of the variability. For this set of variables, relationships with discharge and rainfall appear to be curved with a decreasing relationship evident at higher values indicating a dilution effect. Relationships with the animal covariates are generally positive, especially for larger counts, indicating increases in measured OP with increasing animal waste. Higher grade agricultural land appears to contribute more to OP levels than lower quality grazing land with medium/heavy soils contributing more than chalk or light soils.

Model	Adjusted R^2	Number of covariates
1	70.1%	26
2	65.7%	23

Table 1: Adjusted R^2 values for the Severn LHA

4 Concluding Remarks

Trends and seasonality have been explored in all LHAs in England & Wales for OP, TON and Total Nitrogen with covariate information incorporated for a subset of these LHAs and a space-time interaction incorporated using p-splines for one example area. Future work will include exploring alternative approaches to dealing with large data dimensions and the hierarchical nature of the data.

References

Bowman, A.W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. OUP, Oxford.

Helsel, D.R. (2005). *Nondetects and data analysis: statistics for censored environmental data*. New York: Wiley-Interscience.

Wood, S.N. (2006). *Generalized Additive Models, An Introduction with* `R`. Chapman & Hall, Boca Raton.

Definition of type-specific reference conditions in Mediterranean lagoons[1]

Alberto Basset
Department of Biological and Environmental Sciences and Technologies, University of Salento – 7300 Lecce Italy – alberto.basset@unisalento.it

Enrico Barbone
ARPA Puglia, Corso Trieste 27 – 80100 Bari Italy

Ilaria Rosati
Department of Biological and Environmental Sciences and Technologies, University of Salento – 7300 Lecce Italy

Abstract: Defining 'reference conditions' (sensu Water Framework Directive) in Mediterranean lagoons is a challenging issue since the Mediterranean societies have used lagoons for centuries, lagoons are naturally enriched ecosystems, physically stressed and characterised by strong and unstable internal gradients and lagoons show an high taxonomic redundancy and a low taxonomic similarity. Here, accounting for these peculiarities, we have compared *a priori* and *a posteriori* approaches to identify the main sources of uncertainty in the ecological status of Mediterranean lagoons. Mixed model analysis showed that the *a posteriori* approach emphasises metric-specific ecosystem types and reduces the uncertainty of the ecological status classification when compared with the *a priori* approach based on fixed ecosystem Typology.

Keywords: macroinvertebrate, lagoon, reference conditions, typology, ecological status, mixed models, uncertainty.

1. Introduction

The Water Framework Directive (hereafter WFD) requires EU Member States to classify the ecological status of every water body in Europe larger than some minimum threshold defined in the Directive (WFD, 2000). Ecological status is an ecosystem property, which is a measure of ecosystem functioning and is assumed to be high in aquatic ecosystems totally or nearly totally undisturbed by human activities. Therefore, ecological status of ecosystems is conceptually independent of the natural variability of its structural components, which can be very large conditions on spatial and temporal scales both among and within aquatic ecosystems. Since ecological status of ecosystems is commonly assessed from the characteristics of their biotic components, the natural variability of plant and animal guild attributes, depending on the abiotic context (i.e.,

[1] WISER (Water bodies in Europe: Integrative 530 Systems to assess Ecological status and Recovery) funded by the European Union under the 7th 531 Framework Programme, Theme 6 (Environment including Climate Change) 532 (contract No. 226273), www.wiser.eu.

the niche of the environment, Emlen, 1973; Zobel, 1997), represents a major source of uncertainty in the process of ecological status assessment.
The WFD addresses the uncertainty derived from natural variability with the classification of ecosystems into 'types', defined according to the main drivers of variation of the biotic components using a discrete scale. For example, Typology of lagoon ecosystems in the Mediterranean ecoregion was proposed to be based on a hierarchical organization of three drivers, tidal range, lagoon surface area and water salinity, producing globally 20 types (2 classes of tidal range x 2 classes of surface area x 5 classes of water salinity; Basset et al., 2006; Lucena-Moya et al., 2009). Recently, the degree of confinement was also proposed as a major driver of biotic component variation in Mediterranean lagoons. A classification of ecosystems into types, if the proper drivers are selected, actually reduces the natural variability within every type, but being generally based on the assumption of linear responses along the driver gradients it can incorporate both redundancy between types and uncertainty within a few types. Recently, the application of a mixed model approach was found very effective in optimizing the definition of 'ecosystem types' and assessment of ecological status in Mediterranean lagoons (Barbone et al., 2011), but this *a posteriori* approach may be biased by the data source used and less general than the *a priori* approach.
Here, we have compared the *a priori* and the *a posteriori* approach using a data-set on benthic macroinvertebrate guilds of Mediterranean and Black Sea lagoons and the *a priori* typological classification of Mediterranean and Black Sea lagoons available in the literature or on official documents of the Committee in charge of implementing the methodological procedures of aquatic ecosystem assessment in Europe according to the WFD.

2. Materials and Methods

Data analysis was performed of biotic and abiotic data available at the Transitional Water Platform (www.circlemednet.unisalento.it). Data were originally collected on fourteen Mediterranean and Black Sea lagoon ecosystems in the framework of the European project TWReferenceNet. The studied ecosystems or ecosystem areas were selected because of their high degree of naturality, when compared with the average conditions in the EcoRegional area; all studied ecosystem areas were exposed to low anthropogenic pressures (Table 1) and utilised as potential reference conditions in order to explore the influence of spatial and temporal sources of natural variability. The data used for this study are based on a nested sampling with habitat types (2/3), sites (2) and replicates (5) nested within lagoons (14) and times (2). Abiotic data include measures of pressures, ecosystem physiography and hydrology, at the ecosystem level, and measures of chemical-physical water parameters, at the level of sampling sites/times. Biotic data refer to the macroinvertebrate guilds of the studied lagoons/lagoon areas and include measurements of species composition, numerical abundance and individual traits at the replicate level; data were then aggregated for the analysis at the site level. As individual trait, individual body size was quantified on all sampled individuals as body length and ash free body mass; body mass was not determined in less than 5% for technical problems.
Simple metrics and multi-metric indices were computed from the original data; the former includes measures of species composition and richness, numerical abundance,

diversity, average individual mass and size spectra components, the latter include four main multi-metric indices, namely BAT (Benthic Assessment Tool), BITS (Benthic Index based on Taxonomic Sufficiency), ISS (Index of Size Spectra), M-AMBI (multivariate AMBI).
The amount of variation of both simple and multi-metric indices explained by standard lagoon typologies, based on water salinity, lagoon surface area, tidal range and confinement, or by a posteriori assessed typology with the use of mixed model approaches have been quantified and compared. The uncertainty of ecological status assessments at the level of sampling sites, or lagoons following the two different approaches and the different indices was also addressed.

Table 1. Pressure evaluation on the list of the transitional water ecosystems considered: A = organic load; B = nutrient load; C = hazard substances; D[a] = fishing; E = alien species; F = navigation; G= physical modification; H= average pressure; I[a] = net pressure The intensity of every pressure type was evaluated using a scale of value ranging from 0 (absent) to 4 (4=high).

Transitional waters	Pressures								
	A	B	C	D*	E	F	G	**H**	**I***
Agiasma	2	2	1	3	-	-	2	**2.0**	**1.7**
Logarou	1	2	1	4	-	-	3	**2.2**	**1.7**
Alimini	1	1	-	3	1	-	1	**1.4**	**1.0**
Grado Marano[b]	2	2	3	-	2	3	2	**2.3**	**2.3**
Grado Valle Cavanata	1	1	1	-	1	-	1	**1.0**	**1.0**
Grado Valli da Pesca[b]	1	1	-	4	4	-	2	**2.4**	**2.0**
Le Cesine	-	1	-	-	-	-	2	**1.5**	**1.5**
Margherita di Savoia[b]	2	2	-	1	-	-	4	**2.2**	**2.6**
Torre Guaceto	-	1	1	-	-	-	1	**1.0**	**1.0**
Karavasta	1	1	-	4	1	1	3	**1.8**	**1.4**
Narta[b]	2	2	-	3	-	-	4	**2.7**	**2.6**
Patok	-	-	-	2	1	-	1	**1.3**	**1.0**
Lehaova	1	1	-	1	1	-	1	**1.0**	**1.0**
Sinoe	1	1	-	3		-	3	**2.0**	**1.6**

[a] Fishing pressures are not considered so effective on benthic macro-invertebrates. Net Pressures exclude fishing [b]Pressures at less perturbed sites can be estimated 0.33 of Net pressures.

3. Results

The main results achieved from with the data analysis are listed below:

1. Simple metrics are sensitive to internal lagoon patchiness while multi-metric indices are not;
2. All multi-metric indices showed a significant variability among lagoon ecosystems;
3. Considering only the multi-metric indices, BITS has an higher intrinsic variability than the other three metrics;

4. Water salinity is the typological category accounting for most variability of M-AMBI, BAT and ISS, while confinement was an important source of BITS variability;
5. The *a posteriori* mixed model analysis showed clear metric-specific, type specific reference conditions;
6. Accounting for these reference conditions, the accuracy of both M-AMBI and ISS in the classification of ecological status of the studied lagoons was highly improved (Figure 1);
7. A priori, categorical, lagoon classification also explained part of the multi-metric variation, improving the ecological status classification of the studied lagoons.

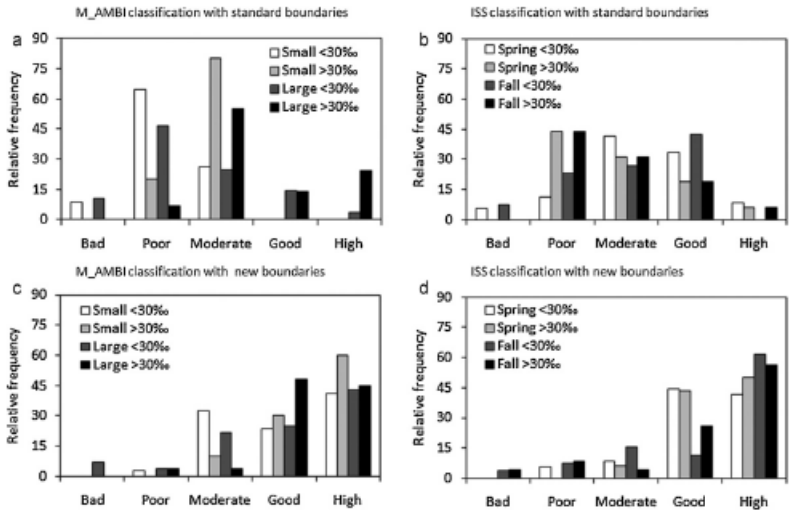


Figure 1: Ecological quality classification of the study sites among type specific categories (da Barbone et al., 2011).

4. Concluding remarks

The comparative analysis of the performance of a priori and a posteriori approaches to the definition of a typological classification of aquatic ecosystem has a major applied implication in the optimization of the regional and national monitoring programs .

References

Barbone, E., Rosati, I., Reizopoulou, S., Basset, A. (2011) Linking classification boundaries to sources of natural variability in transitional waters: a case study of benthic macroinvertebrates. *Ecological Indicators* doi:10.1016/j.ecolind.2011.04.014

Basset, A., Sabetta, L., Fonnesu, A., Mouillot, D., Do Chi, T., Viaroli, P., Giordani, G., Reizopoulou, S., Abbiati, M., Carrada, G.C., (2006). Typology in Mediterranean transitional waters: new challenges and perspectives. *Aquatic Conservation – Marine and Freshwater Ecosystems*, 16, 441–455.

Lucena-Moya, P., Pardo, I., Alvarez, M., (2009). Development of a typology for transitional waters in the Mediterranean ecoregion: The case of the islands. *Estuarine and Coastal Shelf Science*, 82, 61-72.

Ensemble forecasting: status and perspectives

F. Nerozzi, T. Diomede, C. Marsigli, A. Montani, T. Paccagnella
Servizio Idro–Meteo–Clima, ARPA Emilia–Romagna, Italy (fnerozzi@arpa.emr.it)

Abstract: One of the main challenges for Numerical Weather Prediction is the Quantitative Precipitation Forecasting (QPF). The accurate forecast of high-impact weather still remains difficult beyond day 2 and many limited-area ensemble prediction systems have been recently developed so as to provide more reliable forecasts than achievable with a single deterministic forecast. As a consequence the calibration of ensemble precipitation forecasts has become a very demanding task, for improving the QPF, especially as an input to hydrological models. Different calibration techniques are compared: cumulative distribution function, linear regression and analogues method.

Keywords: Ensemble Prediction Systems, Quantitative Precipitation Forecasts, Calibration techniques.

1 Introduction

The first approach to the probabilistic predictions in meteorology occurred in the early seventies, the emphasis being on the study of stochastic–dynamics equations. Recently, other approaches have been developed, one of these is based on the description of the temporal evolution in the phase space of the probabilistic distribution function (PDF) of the model state vector by the Liouville equation (LE), or the Fokker–Plank equation (FPE), if model errors are taken into account through specific random forcing terms in the governing model equations (Ehrendorfer, 1994).

Actually, an approach based on LE and FPE is considered impractical in the context of forecasting forecast skill, because the high dimensionality of the state vector of realistic meteorological models and of the associated phase space (Ehrendorfer, 1994) and a finite ensemble of numerical predictions appears to be the only feasible way to predict the evolution of the atmospheric PDF beyond the range in which error growth can be prescribed by linearized dynamics. Two requirements arise: statistics of this finite ensemble should sample correctly the PDF of analysis errors and model trajectories in the phase space should be good approximations of the corresponding trajectories of the atmosphere (Molteni et al., 1996).

The idea of probabilistic weather predictions is widely accepted now: since 1992, both the National Center for Environmental Prediction (NCEP) and the European Center for Medium–Range Weather Forecast (ECMWF) have been providing weather ensemble predictions (Tracton and Kalnay, 1993 and Palmer et al., 1993).

Specifically, the ECWMF Ensemble Prediction System (EPS) has been tuned for predictions ranging from day 2 to day 15, and is based on a configuration with 50 perturbed and 1 unperturbed (i.e. starting from the ECMWF analysis, say control) members. Perturbed analyses are obtained by adding and subtracting to the operational analysis 25 orthogonal perturbations (representing the observation errors), obtained using a combination of singular vectors, computed to optimize total energy growth over a 48 hours time interval (Molteni et al., 1996). Model uncertainties are simulated by adding stochastic perturbations to the tendencies due to parameterized physical processes (Buizza et al., 1999).

2 The COSMO–LEPS Ensemble Prediction System

One of the main challenges for numerical weather prediction (NWP) is still recognized as quantitative precipitation forecasting. Computer power resources have greatly increased in the last years, thus allowing the generation of more and more sophisticated NWP models with accurate parametrization of physical processes supported by high horizontal and vertical resolution. Nevertheless, the accurate forecast of high-impact weather still remains difficult beyond day 2 and sometimes, also for shorter ranges (Tibaldi et al., 2006).

Many limited-area ensemble prediction systems have been recently developed, either in research or in operational mode, so as to address the need of detailing high-impact weather forecasts at higher and higher resolution and to provide more reliable forecasts than achievable with a single deterministic forecast. The methodology aims at combining the advantages of the probabilistic approach by global ensemble systems with the high-resolution details gained in the mesoscale integrations (Montani et al., 2011).

As far as operational implementations are concerned, the COnsortium for Small-Scale MOdelling Limited-area Ensemble Prediction System (COSMO-LEPS) is based on 16 integrations of the non–hydrostatic mesoscale model COSMO (Montani et al., 2011). In the construction of COSMO–LEPS, an algorithm selects a number of members from the ECMWF ensemble system (Marsigli et al., 2001; Molteni et al., 2001), which are used to provide both initial and boundary conditions to the integrations with the COSMO model.

3 Calibration of the Quantitative Precipitation Forecast

The calibration of precipitation forecast at high resolution is a challenging and quite new scientific issue (Hamill et al. 2008).

Fundel et al. (2009) experienced with reforecast of COSMO-LEPS (30 years) the calibration of the COSMO–LEPS precipitation over Switzerland. They carried out some sensitivity studies in order to determine the impact of the length of the reforecast period. Diomede et al. (2010), focusing the calibration work on the statistical adjustment of 24–h Quantitative Precipitation Forecasts provided by COSMO-LEPS over the Emilia–Romagna region (Northern Italy), have been used the reforecasts run by MeteoSwiss for comparing three calibration techniques: cumulative distribution function, linear regression and analogues method, based on the similarity of the forecasted precipitation fields. Two different implementations of these techniques with respect to the method used for spatial aggregation of the model grid points have been tested: calibrating functions defined either for each model grid point or for eight areas partitioning the Emilia–Romagna region. The calibration process provided a slight improvement for the reliability and skill of the COSMO-LEPS QPFs, except for the autumn season. Generally, the raw and calibrated forecasts were overconfident. Forecasts of lower precipitation events were more skilful than forecasts of higher precipitation events. The calibration functions defined for each model grid point showed higher performance. The lack of improvement related to the CDF and LR–based methods can be ascribed to the lack of a strong relationship between forecast and observed data. Results suggested that weather–regime specific correction functions should be required for improving the COSMO-LEPS QPFs.

4 Concluding remarks

It is expected the calibration of QPF could improve the skill of COSMO-LEPS forecasts, making the system more reliable and the calibrated QPF introduced as an input to hydrological models. In the future, an increase of the horizontal resolution of COSMO-LEPS will be tested. The higher resolution will likely provide more detailed forecasts for the interaction of the flow with orography and will describe with a higher degree of accuracy mesoscale–related processes and local effects. This would have a positive impact on the prediction of a number of those surface fields still nowadays strongly influenced by local effects and not always properly represented in terms of their uncertainty by mesoscale ensemble systems.

References

Buizza R., Miller M. and Palmer T. N. (1999) Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System, ECMWF Technical Memorandum, 279, 26pp.

Diomede T., Marsigli C., Montani A. and Paccagnella T. (2010) Comparison of calibration techniques for a limited-area ensemble precipitation forecast using

reforecasts. In: Proceedings of the Third WMO International Conference On QPE/QPF and Hydrology, Nanjing, China, 1822 October 2010.

Ehrendorfer M. (1994) The Liouville Equation and its potential usefulness for the prediction of forecast skill. Part I: Theory, Mon. Wea. Rev., 122, 703–713.

Fundel F., Walser A., Liniger M. A., Frei, C. and Appenzeller C. (2009) Calibrated precipitation forecasts for a limited-area ensemble forecast system using reforecasts. Mon. Wea. Rev. 138, 176189.

Hamill T. M., Hagedorn R. and Whitaker J.S. (2008) Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts, Mon. Wea. Rev., 136, 2620,2632.

Marsigli C., Montani A., Nerozzi F., Paccagnella T., Tibaldi S., Molteni F., Buizza R., (2001) A strategy for high–resolution ensemble prediction. Part II: Limited–area experiments in four Alpine flood events. Q. J. R. Meteorol. Soc., 127, .

Molteni F., Buizza R., Palmer T.N., Petroliagis T. (1996) The ECMWF Ensemble Prediction System: Methodology and validation. Q. J. R. Meteorol. Soc. 122, 73–119.

Molteni F., Buizza R., Marsigli C., Montani A., Nerozzi F., Paccagnella T. (2001) A strategy for high–resolution ensemble prediction. Part I: Definition of Representative Members and Global Model Experiments. Q. J. R. Meteorol. Soc. 127, 2069–2094.

Montani A., Cesari D., Marsigli C. and Paccagnella T. (2011) Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: main achievements and open challenges, Tellus, 63A, 605–624.

Palmer T. N., Molteni F., Mureau R. and Buizza R. (1993) Ensemble Prediction, ECMWF Seminar Proceedings "Validation of models over Europe": Vol. 1, ECMWF, Shinfield PArk, Reading, RG2–9AX, UK, pp. 21–66.

Tracton M. S. and Kalnay E. (1993), Operational Ensemble Prediction at the National Meteorological Center: Practical Aspects, NMC Notes, 8, 379–398.

Statistical postprocessing for ensembles of numerical weather prediction models

Tilmann Gneiting
University of Heidelberg, t.gneiting@uni-heidelberg.de

Abstract: The past fifteen years have witnessed a radical change in the practice of weather forecasting, in that ensemble prediction systems have been implemented operationally. An ensemble forecast comprises multiple runs of numerical weather prediction models, which differ in initial and lateral boundary conditions, and/or the parameterized representation of physical processes. However, ensemble forecasts are subject to biases and dispersion errors, and thus statistical postprocessing is required, with Bayesian model averaging and ensemble model output statistics being state of the art approaches. Future work is called for to ensure that the postprocessed forecast fields show physically realistic and coherent joint dependence structures across meteorological variables, geographic space and look-ahead times.

Keywords: Bayesian model averaging; ensemble model output statistics; numerical weather prediction; statistical postprocessing

1 Introduction

A major human desire is to make forecasts for an uncertain future. Consequently, forecasts ought to be probabilistic in nature, taking the form of probability distributions over future quantities or events (Dawid 1984; Gneiting 2008). That said, weather forecasting has traditionally been viewed as a deterministic exercise, drawing on highly sophisticated numerical models of the atmosphere. The advent of ensemble prediction systems in the 1990s marks a radical change (Palmer 2002; Gneiting and Raftery 2005). An ensemble forecast comprises multiple runs of numerical weather prediction models, which differ in initial conditions, lateral boundary conditions, and/or the parameterized representation of the atmosphere being used. An example from the University of Washington Mesoscale Ensemble (Grimit and Mass 2002) over Western North America and the Northeast Pacific Ocean is shown in Figure 1.

2 Statistical postprocessing of ensemble weather forecasts

Realizing the full potential of an ensemble forecast requires statistical postprocessing of the model output, to address model biases and dispersion errors.

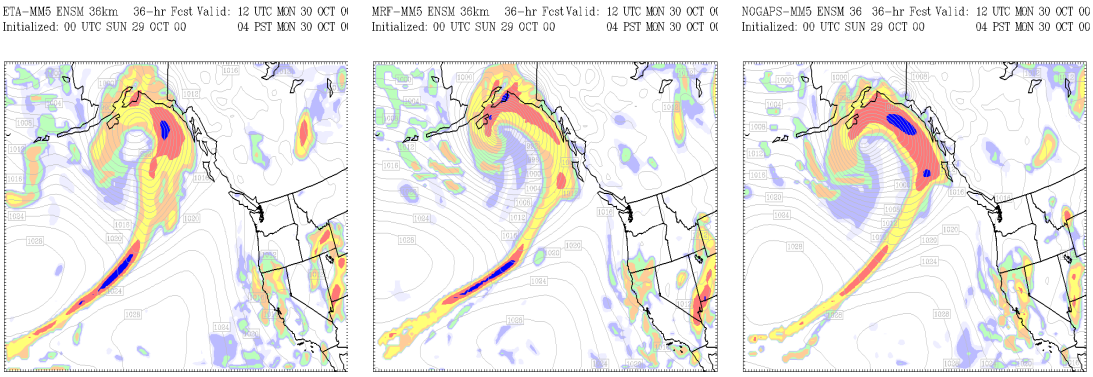


Figure 1: 36-hour ahead ensemble forecast valid October 30, 2000 over Western North America and the Northeast Pacific Ocean, with color representing precipitation amounts. Three members of the University of Washington Mesoscale Ensemble (Grimit and Mass 2002) are shown.

Popular approaches for doing this include the Bayesian model averaging (BMA) method developed by Raftery et al. (2005) and the ensemble model output statistics (EMOS), or heterogeneous regression, technique introduced by Gneiting et al. (2005). The BMA approach employs a mixture distribution, where each mixture component is a parametric probability density associated with an individual ensemble member, with the mixture weight reflecting the member's relative contributions to predictive skill over a training period. In contrast, the EMOS predictive distribution is a single parametric distribution.

To fix the idea, consider an ensemble of NWP forecasts, $f_1, \ldots, f_k$, for temperature, x, at a given time and location. Let $\phi(x; \mu, \sigma^2)$ denote the normal density with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ evaluated at $x \in \mathbb{R}$. The BMA approach of Raftery et al. (2005) employs Gaussian components with a linearly bias-corrected mean. The BMA predictive density for temperature then becomes

$$p(x \,|\, f_1, \ldots, f_k) \;=\; \sum_{i=1}^{k} w_i \, \phi(x; a_i + b_i f_i, \, \sigma^2),$$

with BMA weights, $w_1, \ldots, w_k$, that are nonnegative and sum to 1, bias parameters $a_1, \ldots, a_k$ and $b_1, \ldots, b_k$, and a common variance parameter, σ^2, all of which being estimated from training data over a rolling training period that consists of the recent past. The EMOS approach of Gneiting et al. (2005) employs a single Gaussian predictive density, in that

$$p(x \,|\, f_1, \ldots, f_k) = \phi(x; a + b_1 f_1 + \cdots + b_k f_k, \, c + d s^2),$$

with regression parameters a and $b_1, \ldots, b_k$, and spread parameters c and d, where s^2 is the variance of the ensemble values. The EMOS technique thus is more parsimonious, and the BMA method is more flexible.

While the original methodological development of Raftery et al. (2005) and Gneiting et al. (2005) was addressed at temperature and surface pressure, more recent work aims at the statistical postprocessing of ensemble forecasts for quantitative precipitation (Sloughter et al. 2007), wind speed (Sloughter et al. 2010; Thorarinsdottir and Gneiting 2010) and wind direction (Bao et al. 2010). For a fully Bayesian alternative to the BMA approach of Raftery et al. (2005), see Di Narzio and Cocchi (2010).

3 Challenges for future work

Even though Bayesian model averaging and ensemble model output statistics are state of the art methods, they treat distinct weather variables at distinct geographic locations and distinct look-ahead times independently of each other. This conflicts with key applications such as air traffic control, flood management or winter road maintenance, where it is critically important that the postprocessed forecast fields show physically realistic and coherent joint dependence structures across meteorological variables, geographic space and look-ahead times.

Perhaps the most advanced technique in these directions is the Spatial BMA approach of Berrocal, Raftery and Gneiting (2007), who merged the traditional BMA approach of Raftery et al. (2005) with the geostatistical output perturbation (GOP) technique of Gel, Raftery and Gneiting (2004) to obtain probabilistic temperature field forecasts that honor the spatial structure of observations. Similarly, the Bernoulli-Gamma BMA approach of Sloughter et al. (2007) could be merged with the two-stage spatial method of Berrocal, Raftery and Gneiting (2008), which uses Gaussian copulas, to yield spatially and/or temporally coherent postprocessed forecast fields for quantitative precipitation. Variants of the Schaake shuffle (Clark et al. 2004) provide nonparametric alternatives. Work along these lines is a critical research need in the statistical postprocessing of ensemble weather forecasts, and there is ample scope for continued methodological development, using nonparametric tools, methods of spatial and spatio-temporal statistics, and/or copula techniques.

References

Bao, L., Gneiting, T., Grimit, E. P., Guttorp, P. and Raftery, A. E. (2010). Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. *Monthly Weather Review*, **138**, 1811–1821.

Berrocal, V. J., Raftery, A. E. and Gneiting, T. (2007). Combining spatial statistical and ensemble information for probabilistic weather forecasting. *Monthly Weather Review*, **135**, 1386–1402.

Berrocal, V. J., Raftery, A. E. and Gneiting, T. (2008). Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Annals of Applied Statistics*, **2**, 1170–1193.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. and Wilby, R. (2004). The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, **5**, 243–262.

Dawid, A. P. (1984). Statistical theory: The prequential approach (with discussion and rejoinder). *Journal of the Royal Statistical Society Series A: General*, **147**, 278–292.

Di Narzo, A. F. and Cocchi, D. (2010). A Bayesian hierarchical approach to ensemble weather forecasting. *Journal of the Royal Statistical Society Series C: Applied Statistics*, **59**, 405–422.

Gel, Y., Raftery, A. E. and Gneiting, T. (2004). Calibrated probabilistic mesoscale weather field forecasting: The Geostatistical Output Perturbation (GOP) method (with discussion). *Journal of the American Statistical Association*, **99**, 575–587.

Gneiting, T. (2008). Editorial: Probabilistic forecasting. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **171**, 319–321.

Gneiting, T. and Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, **310**, 248–249.

Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133**, 1098–1118.

Grimit, E. P. and Mass, C. F. (2002). Initial results of a mesoscale short-range ensemble system over the Pacific Northwest. *Weather and Forecasting*, **17**, 192–205.

Palmer, T. N. (2002). The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society*, **128**, 747–774.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133**, 1155–1174.

Sloughter, J. M., Raftery A. E., Gneiting, T. and Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, **135**, 3209–3220.

Sloughter, J. M., Gneiting, T. and Raftery, A. E. (2010). Probabilistic wind forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, **105**, 25–35.

Thorarinsdottir, T. L. and Gneiting, T. (2010). Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **173**, 371–388.

Generalised Kriging with Environmental Applications

Luigi Ippoliti
DMQTE, University G. d'Annunzio, Chieti-Pescara
email: ippoliti@unich.it

Abstract: We consider the problem of spatial interpolation and outline the theory behind kriging and more specifically intrinsic random function kriging. We also mention thin-plate spline theory and show its link with kriging in order to overcome problems in which the available data are not sufficient to estimate the spatial covariance structure of the process. A generalization of the theory to include kriging with directional derivatives is also considered.

Keywords: Kriging, Spatial processes; Thin-plate splines, Derivative process.

1 Introduction

In this paper we are interested in predicting or interpolating values of a spatial process X. Many models for spatial and spatio-temporal data use Gaussian Random Fields (GRFs), and the geostatistical approach of specifying the covariance function, and hence determining the variance matrix $\mathbf{\Sigma}$. In this paper, the approach is to assume that a specified covariance function is of interest and that interpolation of the process is required. However, we consider the case in which there is only a little prior knowledge of the field so that an estimation of the covariance structure is difficult or unfeasible. In this framework, we exploit the one-to-one-correspondence between Reproducing Kernel Hilbert Spaces (RKHSs) and positive semi-definite (p.s.d) functions and show that thin-plate splines, considered as a special case of RKHSs, provides a useful solution of the interpolation problem.

2 Materials and Methods

2.1 Kriging

In this section we provide a brief introduction of the kriging predictor for both stationary and intrinsic random fields. For convenience, for known results we mainly refer here to Cressie (1993) and Mardia *et al.* (1996).
Suppose that a spatial process, $\{X(\mathbf{t}), \mathbf{t} \in \mathcal{R}^d\}$, is observed at sites $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n$ with $\mathbf{t}_i = (t_i[1], \ldots, t_i[d])^T$. An important problem in spatial analysis is to predict $X(\mathbf{t}_0)$ at some new site $\mathbf{t}_0 \in \mathcal{R}^d$. The problem reduces to find a predictor of the form

$$\hat{X}(\mathbf{t}_0) = \boldsymbol{\lambda}^T \mathbf{x} \tag{1}$$

where $\mathbf{x} = [x(\mathbf{t}_1), x(\mathbf{t}_2), \ldots, x(\mathbf{t}_n)]^T$ and $\boldsymbol{\lambda}$ is a $(n, 1)$ coefficient vector chosen to minimise the prediction variance, $E\left\{[X(\mathbf{t}_0) - \hat{X}(\mathbf{t}_0)]^2\right\}$, subject to the unbiasedness constraint, $E\left[X(\mathbf{t}_0) - \hat{X}(\mathbf{t}_0)\right] = 0$. To provide a solution at this problem, suppose the random field has some polynomial drift of order r. Also let $\mathcal{G}_r$ be the space of these polynomial terms, whose dimension is given by $M = \begin{pmatrix} d+r \\ d \end{pmatrix}$. Denote with $\mathbf{U} = \{u_{im}\};\ u_{im} = \mathbf{t}_i^m, i = 1, 2, \ldots, n, |m| \leq r$, the (n, M) drift (design) matrix and with $\mathbf{u}_0$ the vector of drift terms at $\mathbf{t}_0$, with elements $\mathbf{t}_0^m, |m| \leq r$. Assume also that $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$ is a non-singular covariance matrix with entries obtained by defining a "potential" function, $\sigma(|\mathbf{h}|) = \sigma(|\mathbf{t}_i - \mathbf{t}_j|),\ i, j = 1, \ldots, n$. Finally, let $\sigma^2 = \sigma(0)$ and $\boldsymbol{\sigma}_0$ a covariance vector with elements $\sigma(|\mathbf{t}_0 - \mathbf{t}_i|),\ i = 1, \ldots, n$. Following this notation, it can be shown that $\boldsymbol{\lambda}$ is given by

$$\boldsymbol{\lambda} = \mathbf{A}\mathbf{u}_0 + \mathbf{B}\boldsymbol{\sigma}_0 \tag{2}$$

where $\mathbf{A}$ and $\mathbf{B}$ are (n, M) and (n, n) matrices respectively, whose form depends on the underlying assumptions about the random process, and in particular on the properties of the covariance matrix $\boldsymbol{\Sigma}$. Also, note that for the case when $X(\mathbf{t})$ is an intrinsic random field, we have the additional constraint that the coefficients of the prediction error, $\left[\sum_i \lambda_i x(\mathbf{t}_i) - X(\mathbf{t}_0)\right]$, are generalised increments of order r. This constraint can be written in the form $\mathbf{U}^T\boldsymbol{\lambda} = \mathbf{u}_0$, which is the only constraint we need in the kriging problem (Cressie, 1993).

In the stationary random field case, it is known that $\boldsymbol{\Sigma}$ is positive definite. In the case of an intrinsic random field, the assumption that $\boldsymbol{\Sigma}$ is positive definite is no longer valid. Hence we must find an alternative form for the kriging predictor which only requires $\boldsymbol{\Sigma}$ to be conditionally positive definite. It can be shown that $\boldsymbol{\lambda}$ still takes the form of equation (2), but $\mathbf{A}$ and $\mathbf{B}$ must be represented in a way which does not require $\boldsymbol{\Sigma}$ to be positive definite. Provided $\boldsymbol{\Sigma}$ is non-singular, one method for determining $\mathbf{A}$ and $\mathbf{B}$ is to define the matrices (Mardia *et al.*, 1996)

$$\mathbf{K} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{U} \\ \mathbf{U}^T & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \mathbf{K}^{-1} = \begin{bmatrix} \mathbf{K}^{11} & \mathbf{K}^{12} \\ \mathbf{K}^{21} & \mathbf{K}^{22} \end{bmatrix}$$

such that

$$\mathbf{A} = \mathbf{K}^{12} \quad \text{and} \quad \mathbf{B} = \mathbf{K}^{11}.$$

Then, by setting $\mathbf{a} = \mathbf{A}^T\mathbf{x}$ and $\mathbf{b} = \mathbf{B}^T\mathbf{x}$, we may write

$$\begin{aligned}\hat{X}(\mathbf{t}_0) &= \mathbf{a}^T\mathbf{u}_0 + \mathbf{b}^T\boldsymbol{\sigma}_0 \\ &= \sum_{j=1}^{M} a_j u_{j0} + \sum_{i=1}^{n} b_j \sigma(\mathbf{t}_0, \mathbf{t}_i). \end{aligned} \tag{3}$$

One possible common choice for $\sigma(\cdot)$ is any valid covariance function for a stationary stochastic process in space, for which any null space of functions $\mathcal{G}_r$ will suffice. However, in all cases in which the number of spatial sites is small and the estimation of the covariance structure appears difficult, the following class of functions is useful:

$$\sigma_\alpha(\mathbf{h}) = \begin{cases} |\mathbf{h}|^{2\alpha} \log |\mathbf{h}|, & \text{for } \alpha \text{ an integer} \\ |\mathbf{h}|^{2\alpha}, & \text{for } \alpha \text{ not an integer} \end{cases}$$

where α is a smoothness parameter which can be specified ahead of time. Note that this class of functions represents the covariance functions for intrinsic random functions of order $k = [\alpha]$, with $[\alpha]$ the integer part of α. Also, these functions are self-similar and so $\sigma_\alpha(\mathbf{h})$ and $\sigma_\alpha(c\mathbf{h})$ yield the same predictions.

2.2 Kriging with derivative information

There are some cases of interest in which the information about objects (e.g. plants and weeds in crop images) comes from the boundary, which is a continuous curve. In this framework, the kriging predictor can be used in order to modelling the continuous outline of the object. A key aspect to our particular problem is thus the introduction of some extra information, such as derivatives, in order to get a better performance of the modelling procedure.

Suppose that all the known values and derivatives of a specific spatial pattern are collected into a vector $\mathbf{y}$. Let $\boldsymbol{\kappa}$ be a vector of corresponding indices to show the order of the derivative; for example, assuming $d = 1$, $\kappa_i = 0$ if y_i is a data value, $\kappa_i = 1$ if y_i is a first derivative. For each site $\mathbf{t}_j$ there may be several choices of κ_i if the value of the function and of some of its derivatives are all known at that site. The problem is, as before, to find a coefficient vector $\boldsymbol{\lambda}$ such that $\hat{X}(\mathbf{t}_0) = \boldsymbol{\lambda}^T\mathbf{y}$ is the best unbiased linear estimator of $X(\mathbf{t}_0)$.

Let $\kappa_i = \big(\kappa_i[1], \ldots, \kappa_i[d]\big)$ be a $d-$dimensional multi-index of non-negative integers with $|\kappa_i| = \kappa_i[1] + \ldots + \kappa_i[d]$. If we have derivative information of order $p = |\kappa_i|$, then α must satisfies the inequality $\alpha > p$. In the following we also take the smallest drift allowable, that is $r = [\alpha]$. In order to define the $d-$dimensional kriging predictor for derivatives of order up to and including p and with polynomial drift of order r, the covariance matrix, $\boldsymbol{\Sigma}$, consists of entries which have the form

$$\sigma_{ij} = (-1)^{|\kappa_j|}\sigma_\alpha^{(\kappa_i+\kappa_j)}(\mathbf{t}_i - \mathbf{t}_j), \quad 1 \le i, j \le n$$

where $\sigma_\alpha^{(\kappa_i)}$ denotes the partial derivative of $\sigma_\alpha(\mathbf{h})$ of order κ_i. The drift matrix $\mathbf{U}$ instead consists of elements

$$u_{im} = \frac{\partial^{|\kappa_i|}}{\partial \mathbf{t}_i^{\kappa_i}}(\mathbf{t}_i^m), \quad 1 \leq i \leq n, \quad |m| \leq r.$$

3 Applications

In the following we provide a list of applications in which the kriging framework, as outlined in sections 2.1 and 2.2, plays an important role.
Spatial and spatio-temporal data occur widely. There is often interest in predicting or interpolating values. Some data sets may have missing values at some sites, or at some times. These missing values may be separated or clumped. Separated missing values may, for example, occur because of random instrument malfunctions. An example of clumped missing values occurs with passive satellite images when there is cloud cover.
A different application when prediction or interpolation is required is checking observations which may appear to be aberrant or influential. Typical influence and outlier statistics are based on estimating the values using all other values.
A wide range of methods used for constructing optimal spatial sampling designs are also based on sampling schemes with minimal prediction variance. However, a basic problem of this approach is that before the actual sampling takes place there is only a little prior knowledge of the field. To overcome this problem, we may set $\alpha = 1$ and chose $\sigma(\mathbf{h}) = |\mathbf{h}|^2 \log |\mathbf{h}|$, which is conditionally positive definite whenever the null space contains the linear trend, *i.e.* $\mathcal{G}_r = span\,(1, t[1], t[2])\,, \mathbf{t} \in \mathcal{R}^2$. In this case, it turns out that the kriging function (1) is an interpolating thin-plate spline.
Finally, we note that the appearance of agricultural products can be evaluated by considering their size, shape, form and the absence of visual defects. Among these features the shape, also measured through the outline of the object, plays a crucial role. Description of agricultural product shape is often necessary in research fields for a range of different purposes, including the investigation of shape for cultivar descriptions, plant variety or cultivar patents and evaluation of consumer decision performance.

References

Cressie, N. (1993). *Statistics for Spatial Data*. Revised edition. New York: Wiley.

Mardia K.V., Kent, J. T., Goodall C.R., Little, J.A. (1998). Kriging and Splines with Derivative Information. *Biometrika*, 83, 217-285

Variograms to Guide Spatial Sampling for Kriging

Margaret A. Oliver, Soil Research Centre, University of Reading, Reading, RG6 6DW United Kingdom. m.a.oliver@reading.ac.uk
Ruth Kerry, Brigham Young University, Provo, Utah, USA.

Abstract: Detailed information on soil to manage polluted or agricultural sites is often prohibitively expensive to obtain. To sample adequately, the approximate scale of spatial variation needs to be known. If soil data are available, variograms can be computed and used to determine the kriging errors for several grid intervals and an interval selected to meet a specific tolerable error. In the absence of prior knowledge, if soil properties appear related to ancillary data such as aerial photographs, elevation or apparent electrical conductivity (EC_a), individual or multivariate variograms of such data may indicate the scale of variation in the soil. If the scale of variation indicates too few data to compute a reliable variogram conventionally, it can be estimated by residual maximum likelihood or a standardized variogram from ancillary data can be used.

Keywords: variogram, residual maximum likelihood (REML), standardized variogram

1. Introduction

Soil properties can vary at markedly different spatial scales within sites of interest, such as fields. The variation comprises that over short distances of a few metres and over longer distances of tens or hundreds of metres. For most environmental and agricultural management, it is variation over tens or hundreds of metres that managers want to resolve and we can regard the short-range variation as 'noise' or a sampling effect. Many soil attributes have to be determined from samples taken in the field, therefore there is a need to predict accurately at places where there are no data. Kriging provides a sound basis for prediction leading to accurate digital mapping for managing soil attributes (Oliver, 2010). The accuracy of kriged and other interpolated predictions, however, depends on the quality of sample information to compute accurate variograms and availability of spatially dependent data from which to predict (Webster and Oliver, 2007). This means that sampling should be at an interval that is well within the correlation range of spatial variation. Therefore, it is essential that the spatial scales of variation in the properties of most importance for environmental and agricultural management are used to guide sampling.

Sampling on a grid is often used because it provides an even cover of values and minimizes the maximum estimation variance (or error) for a given grid interval and it is efficient for sample collection in the field. If variograms of soil properties from previous surveys exist for an area with a similar soil parent material, they can be used with the kriging equations to determine an optimal grid interval. If the scale of variation is large, the sampling intervals recommended by this method will also be large and there may be too few data from which to compute a reliable variogram by the usual method of moments estimator. Webster and Oliver (1992) showed that at least 100 data are required to compute a reliable variogram in this way from isotropic data. However, Kerry and Oliver (2007) have shown that a variogram estimated by residual maximum

likelihood (REML) can provide more accurate predictions with fewer data than one estimated conventionally. For some soil properties, the variation might be evident in remote and proximally sensed imagery. Variograms computed from such ancillary data can be used to determine the approximate scale of spatial variation. A standardized variogram based on ancillary data or existing variograms of soil properties can also be used to krige spatially dependent (Kerry and Oliver, 2008). We illustrate these methods with a case study in England.

2. Materials and Methods

Matheron's method of moments (MoM) estimator to compute the variogram is given by

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})} \sum_{i=1}^{m(\mathbf{h})} \{z(\mathbf{x}_i) - z(\mathbf{x}_i + \mathbf{h})\}^2, \quad (1)$$

where $z(\mathbf{x}_i)$ and $z(\mathbf{x}_i+\mathbf{h})$ are the actual values of Z at places $\mathbf{x}_i$ and $\mathbf{x}_i+\mathbf{h}$, and $m(\mathbf{h})$ is the number of paired comparisons at lag $\mathbf{h}$. The parameters of the model fitted to the experimental variogram can be used with the data for prediction at points or over blocks. Kriged predictions are a weighted average of the data, $z(\mathbf{x}_1)$, $z(\mathbf{x}_2)$, …, $z(\mathbf{x}_n)$, at the unknown point or block, B,

$$\hat{Z}(B) = \sum_{i=1}^{n} \lambda_i z(\mathbf{x}_i), \quad (2)$$

where n usually represents the data points within the local neighbourhood and λ_i are the weights. To ensure that the estimate is unbiased the weights are made to sum to one. The estimation variance of $\hat{Z}(B)$ is

$$\mathrm{var}\left[\hat{Z}(B)\right] = \mathrm{E}\left[\left\{\hat{Z}(B) - Z(B)\right\}^2\right] = 2\sum_{i=1}^{n} \lambda_i \bar{\gamma}(\mathbf{x}_i, B) - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j) - \bar{\gamma}(B, B), \quad (3)$$

where $\bar{\gamma}(\mathbf{x}_i, B)$ is the average semivariance between data point $\mathbf{x}_i$ and the target block B, and $\bar{\gamma}(B, B)$ is the average semivariance within B, the within block variance. The kriging error is the square root of this.

McBratney et al. (1981) showed how the variogram and kriging equations could be used to determine an optimal sampling interval for prediction by kriging before obtaining new data from a survey. The kriging weights, and also the kriging variances or errors, depend on the configuration of the sampling points in relation to the target point or block and on the variogram and not depend on the observed values at these points. Therefore if we have a variogram function from a previous survey we can determine the kriging errors for any grid size before sampling.

The experimental multivariate variogram **(**Bourgault and Marcotte, 1991) was computed from aerial photograph data by the standard formula adapted for the multivariate case:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})} \sum_{i=1}^{m(\mathbf{h})} \{\mathbf{z}(\mathbf{x}_i) - \mathbf{z}(\mathbf{x}_i + \mathbf{h})\}^{\mathrm{T}} \mathbf{M} \{\mathbf{z}(\mathbf{x}_i) - \mathbf{z}(\mathbf{x}_i + \mathbf{h})\}, \quad (4)$$

where $\mathrm{z}(\mathbf{x}_i)$ and $\mathrm{z}(\mathbf{x}_i + \mathrm{h})$ are the vectors of observations at $\mathbf{x}_i$ and $\mathbf{x}_i + \mathbf{h}$, T is the transpose and $\mathbf{M}$ is a $p \times p$ positive-definite symmetric matrix defining the relations between the variables.

Pardo-Igúzquiza (1998) suggested that a reliable variogram could be computed from a ‘few dozen’ data by maximum likelihood or residual maximum likelihood (REML). Kerry and Oliver (2007) examined this idea further and suggested that 50 to 60 data might suffice (this paper also provides the detail on the theory of the method).

3. Results

The model parameters of a variogram computed from loss on ignition (LOI) data of a field in Wallingford, Oxfordshire, England were used to determine the kriging errors over blocks of various sizes and for a range of grid intervals. The kriging errors are plotted against grid spacing Fig. 1a and a suitable sampling interval would be 120 m for a tolerable error of 0.5%.

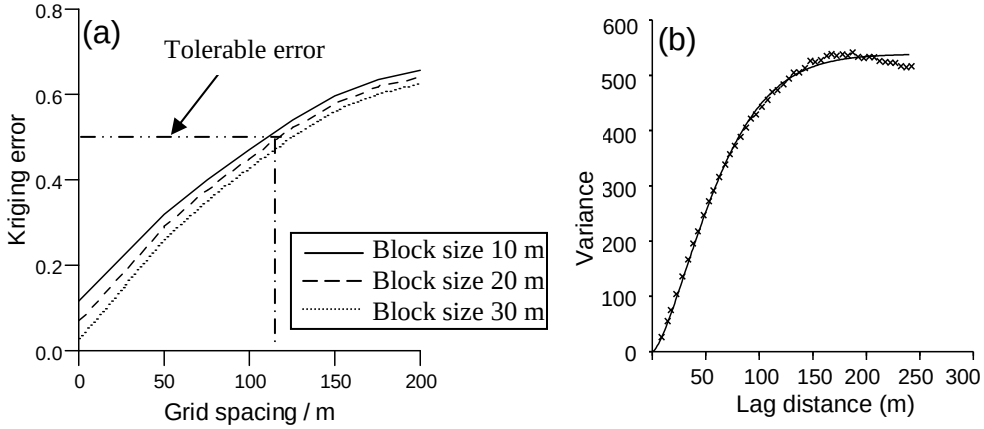


Figure 1: (a) Graph of kriging error against grid spacing for loss on ignition and (b) multivariate variogram of aerial photograph data at Wallingford, Oxfordshire, England.

The multivariate variogram computed from the red, green and blue wavebands of an aerial photograph of bare soil at Wallingford was fitted by a stable exponential function with an approximate range of 205 m (Fig. 1b). Based on less than half the variogram range, this suggested a sampling interval of about 90 m. Figure 2a–c shows kriged maps based on the conventional variograms with the original data on a 30-m grid, the suggested interval of 90 m and for 50 sites based on a 120-m grid with additional samples at 60 m, respectively for LOI at Wallingford. Figure 2d,e shows the kriged maps of LOI based on the 90-m grid with a variogram estimated by REML, and based on a 120-m grid with 15 additional targeted samples estimated by the variogram in Fig. 1b standardized to a sill of unity. Figure 1c–e shows that additional samples at a shorter interval, a variogram estimated by REML or a standardized variogram improve estimates from sparse data.

4. Concluding remarks

The results show the importance of knowing the scale of spatial variation, of having data at distances shorter than half the range of correlation and of alternative methods of estimating the variogram.

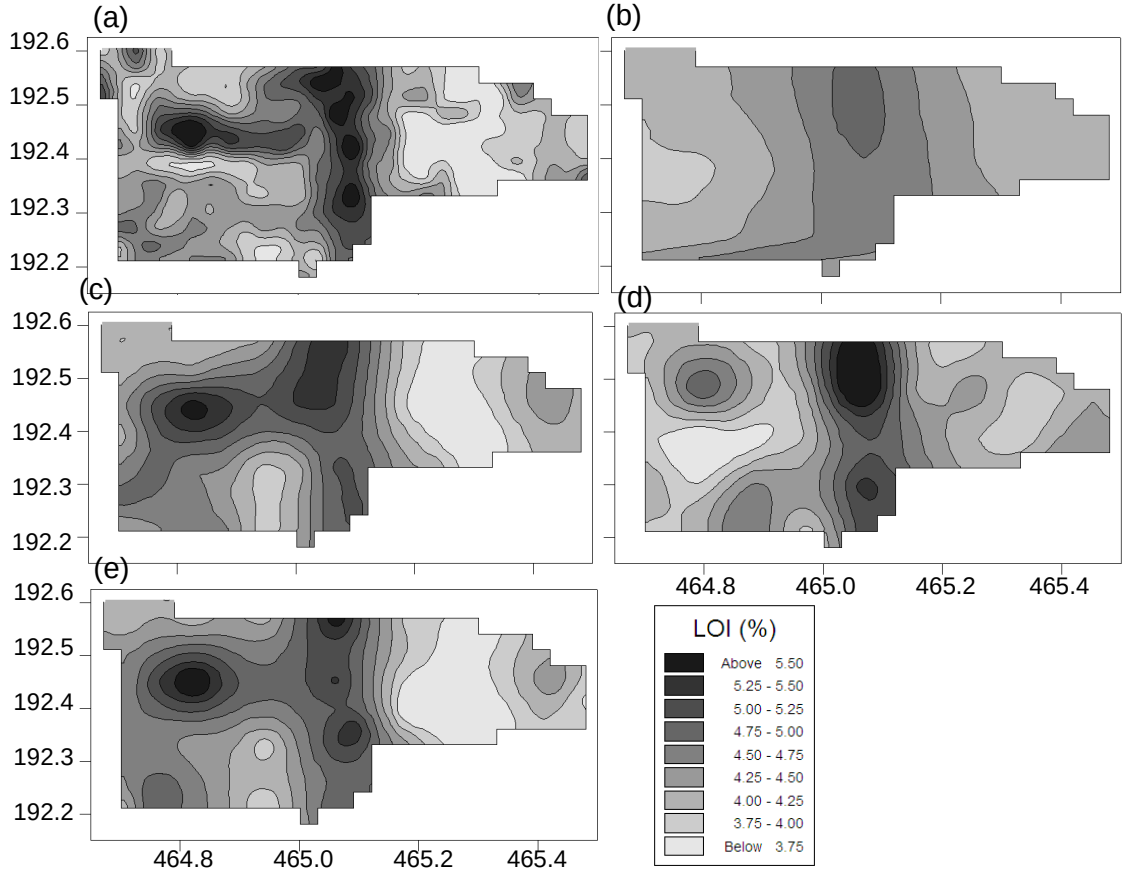


Figure 2: Kriged maps of LOI for Wallingford: with variograms estimated by MoM (a) 30-m grid (296 data), (b) 90-m grid (36 data), (c) 120-m grid + samples at 60-m (50 data); (e) 90-m grid with variogram estimated by REML and (d) 120-m grid + 15 targeted samples with standardized variogram.

References

Bourgault, G., Marcotte, D. (1991) Multivariable variogram and its application to the linear model of coregionalization. *Mathematical Geology*, 23, 899–928.

Kerry, R., Oliver, M. A. (2007) Sampling requirements for variograms of soil properties computed by the method of moments and residual maximum likelihood. *Geoderma*, 140, 383–396.

Kerry, R., Oliver, M. A. (2008) Determining nugget:sill ratios of standardized variograms from aerial photographs to krige sparse soil data. *Precision Agriculture*, 9, 33–56.

McBratney, A. B., Webster, R., Burgess, T. M. (1981) The design of optimal sampling schemes for local estimation and mapping of regionalized variables. I. Theory and method. *Computers & Geosciences*, 7, 331–334.

Oliver, M.A. (2010) An overview of geostatistics and precision agriculture, in: *Geostatistical Applications for Precision Agriculture*, M.A. Oliver (Ed.), Springer, 1–34.

Pardo-Igúzquiza, E. (1998) Maximum likelihood estimation of spatial covariance parameters. *Mathematical Geology*, 30, 95–107.

Webster, R., Oliver, M. A. (1992) Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*, 43, 177–192.

Clustering of environmental functional data

Andrea Pastore and Stefano Tonellato
Università Ca' Foscari Venezia, Department of Economics, stone@unive.it

Roberto Pastres
Università Ca' Foscari Venezia, Department of Environmental Sciences, Informatics and Statistics

Abstract: Often environmental scientists face the problem of clustering different sites, areas or stations in a monitoring network in order to identify some common features among data collected at different locations. In a functional data analysis approach, each location can be seen as a specific individual, on which noisy observations from a continuous random function are collected at discrete times. The definition of suitable models for samples of such functional observations, can provide useful insights about the dynamics of the variables of interest. In such a context, a cluster can be defined as a group of individuals (i.e. locations, stations, areas etc.) where the observed trajectories share common salient features. We present some classification results in a water quality network and focus on some open issues.

Keywords: Cluster analysis, Functional data, Water quality.

1 Introduction

Often environmental scientists face the problem of clustering different sites, areas or stations in a monitoring network in order to identify some common features among data collected at different locations. It is a common practice to use standard classification methods such as k-means or hierarchical classifiers, by considering temporal (e.g. annual) averages of one or more variables measured at each site. This is clearly a limitation, since the whole information about the dynamics of the observed variables is lost. Moreover, such methods do not take into account the uncertainty that should characterise any partition based on sample information. The combination of functional data analysis (Ramsey and Silverman, 2005; Ferraty and Vieu, 2006) and probabilistic cluster analysis methods (Banfield and Raftery, 1993), which allow one to estimate the probability that a given object belongs to a given group, represents, in our opinion, an important step towards a better understanding of environmental data.

Here, we shall provide a classification of the sites of a water quality monitoring network located in Venice Lagoon, by using a trophic index (TRIX, Vollenweider et al. 1998). We apply a classification method based on functional data analysis,

introduced by James and Sugar (2003), which allows to take into account sample information about the temporal dynamics of the variable of interest, as well as quantify the uncertainty in the partition.

2 Classification of functional data

Grossly speaking, functional data analysis methods look at time series of data collected on each individual, in our case on each site, as measurements of a continuous function taken at a finite number of instants and corrupted by noise. Any observed trajectory can be seen as the noisy measurement of an unobservable curve, which is the object of interest. Following the classification method proposed by James and Sugar (2003), data are modelled as a mixture of Gaussian spline regressions, where each mixture represents a model for a specific cluster. Spline coefficients are the sum of a deterministic term, which represents the cluster effect on the mean of the variable, and a stochastic component, which represents an individual (site-specific) random effect. Parameters can be estimated via maximum likelihood. Mixture weights can be seen as prior membership probabilities of any site. The application of Bayes theorem, after plugging maximum likelihood estimates into the model, leads to *posterior* membership probabilities for each site in the network (Banfield and Raftery, 1993). A generic monitoring station is then allocated to the group which encompasses it with highest posterior probability. The number of groups, i.e. the number of mixture components, is selected by using BIC criterion.

3 Site classification in terms of water quality

The data. Venice Lagoon, with an extension of about 500 km^2 is one of the largest wet areas in Europe. It is a shallow water system with average depth of one meter crossed by a network of canals which determine a rather complex hydrodynamic circulation. As other European estuaries and lagoons, it is classified as a transition water body. Overall, the tributary discharge is about 30 m^3/sec. Rivers bring in freshwater, nutrients and pollutants, whereas tides bring in marine water. Internal hydrodynamics disperse the pollutants and, eventually, dissolved compounds are exported to the sea.

Data were collected at 30 monitoring sites which are shown on the map in figure 1. The first chacter of site labels identifies a particular category: letter B means that the site is located in a shallow area, letter C indicates that the site is located on a canal and letter M identifies sites located in the coastal area, next to the Lagoon. The same figure shows (in blue) the network of canals which are very influential in the Lagoon hydrodynamics and must be taken into account when interpreting classification results. Measurements were repeated in time at 38 subsequent instants (in the period ranging from January 16th, 2001 to December 17th, 2003) corresponding to neap tides. We considered a subset of the variables which have been

monitored, namely: chlorophyll-a (CHL-a), dissolved oxygen (DOX), total nitrate (NIT) and reactive phosphorus (PPO4). CHL-a and DOX can be taken as proxies for actual primary production. Even though in shallow lagoons and coastal areas, including the lagoon of Venice, macroalgae and seagrasses usually account for the major fraction of the production, phytoplanctonic production is extremely important, since the planktonic compartment represents a source of food for fish juvenils and shellfish. The concentrations of dissolved oxygen, total nitrate and reactive phosphorus provide information about the trophic potential of a water body. In fact, an excess of these chemicals could enhance the primary production of phytoplankton and macroalgae and cause the symptoms of eutrophication, as happened in Venice Lagoon in the 1970ies an 1980ies.

TRIX. TRIX is a widely used trophic index for marine coastal waters proposed by Vollenweider et al. (1998). It considers both factors that are direct expressions of productivity (chlorophill-a and dissolved oxigen) and nutritional factors (nitrogen and phosphorous). Some alternative formulations have been proposed. Here we consider the following one:

$$TRIX = \frac{\log_{10}(\text{CHL-a} \times \text{DOX} \times \text{NIT} \times \text{PPO4}) + 1.5}{1.2},\ TRIX \in [1, 10]$$

where DOX is the absolute deviation of oxygen from saturation and the other symbols indicate the concentrations, in mg/m^3, of the compounds mentioned above. The values of TRIX range from 1 to 10: low values indicate oligotrophy (scarcity of nutrients); high values indicate hypertrophy (exeedence of nutrients). A water body in a good trophic state should not exceed the value 5.

4 Results

In our application we identified two groups: the first one characterised by good values of TRIX and the second one exhibiting high TRIX values for the most part of the sample period. Figure 1 shows the raw data, the group specific mean trajectories and individual mean trajectories. The same figure shows a map where two spatial clusters are clearly identified. It is worth to note, however, that the posterior membership probabilities of sites $B11$, $C06$, $C01$, and $C05$ range between 0.54 and 0.87, indicating a rather strong uncertainty in their allocation to one of the two groups (for the remaining sites, the allocation probability was always higher than or equal to 0.99).

An explicit treatment of spatial dependence has not yet been developed for the class of models we have considered here. Important advances in this direction have been made in the Bayesian nonparametrics literature and research in this field is under way.

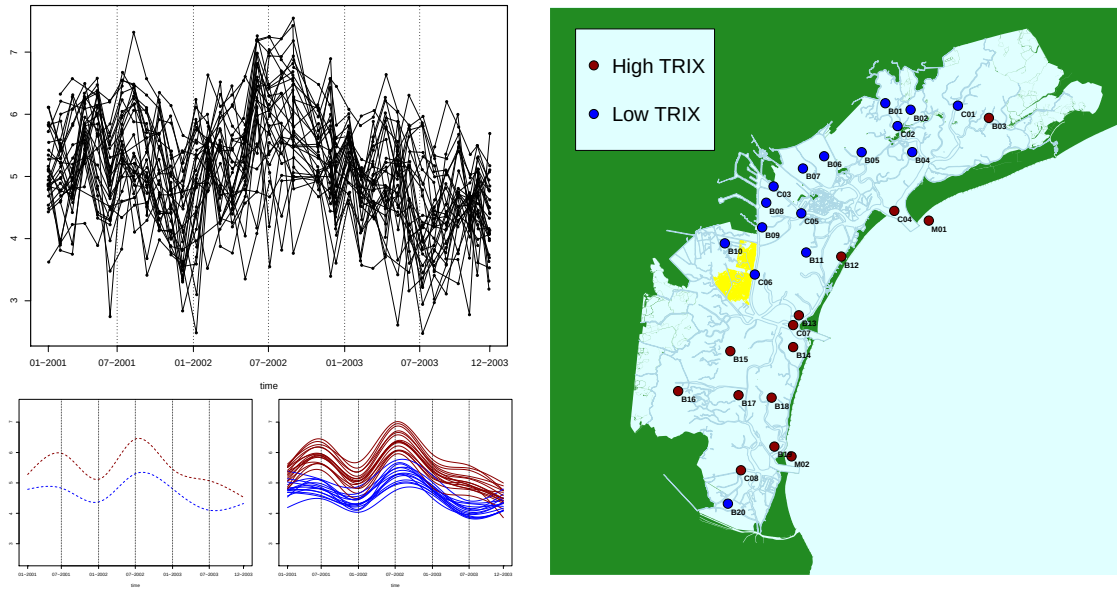


Figure 1: Plots of raw data, group specific mean trajectories, individual mean trajectories and map of monitoring sites (red="high TRIX"; blue="low TRIX").

References

Banfield J. D. and Raftery A. E. (1993) Model-Based Gaussian and Non-Gaussian Clustering, *Biometrics*, 49, 803-821.

Ferraty F. and Vieu P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, Springer, New York.

James G. M. and Sugar C. A. (2003), Clustering for Sparsely Sampled Functional Data, *Journal of the American Statistical Association*, 98, 397-408.

Ramsey J. O. and Silverman B. W. (2005) *Functional Data Analysis*, Springer, New York.

Vollenweider R. A. , Giovanardi F., Montanari G. and Rinaldi A. (1998) Characterization of the trophic conditions of marine coastal waters with special reference to the NW Adriatic Sea: Proposal for a trophic scale, turbidity and generalized water quality index. *Environmetrics*, 9, 329-357.

Spatially correlated functional data[1]

Jorge Mateu
Department of Mathematics, University Jaume I, Castellon, Spain.
mateu@mat.uji.es

Abstract: Observing complete functions as a result of random experiments is nowadays possible by the development of real-time measurement instruments and data storage resources. Functional data analysis deals with the statistical description and modeling of samples of random functions. Functional versions for a wide range of statistical tools have been recently developed. Here we are interested in the case of functional data presenting spatial dependence, and the problem is handled from the geostatistical and point process contexts. Functional kriging prediction and clustering are developed. Additionally, we propose functional global and local marked second-order characteristics.

Keywords: Basis functions, Functional clustering, Functional kriging, LISA functions, Trace-variogram

1 Introduction

In many fields of environmental sciences the observations consist of samples of random functions. Since the early nineties, Functional Data Analysis (FDA) has been used to model this type of data (Ramsay and Dalzell, 1991). From the FDA point of view, each curve corresponds to one observation, that is, the basic unit of information is the entire observed function rather than a string of numbers. Functional versions for many branches of statistics have been given (Ramsay and Silverman, 2005).

The standard statistical techniques for modeling functional data are focused on independent functions. However, in several disciplines of applied sciences there exists an increasing interest in modeling correlated functional data: this is the case when samples of functions are observed over a discrete set of time points (*temporally correlated functional data*) or when these functions are observed in different sites of a region (*spatially correlated functional data*). In these cases some statistical methods for modeling correlated variables have been adapted to the functional context.

We can define a spatial functional process as $\left\{\boldsymbol{\chi}_s, s \in D \subset \mathbb{R}^d\right\}$ where s is a generic data location in the d-dimensional Euclidean space, the set $D \subset \mathbb{R}^d$ can be fixed or random, and $\boldsymbol{\chi}_s$ are functional random variables, defined as random elements taking values in an infinite dimensional space (or functional space). Typically $\boldsymbol{\chi}_s$

[1]Research partially supported by the Spanish Ministry of Education and Science through grant MTM2010-14961, and Bancaja grant P1-1B2008-27.

is a real function from $[a, b] \subset \mathbb{R}$ to $\mathbb{R}$. The nature of the set D allows to classify spatial functional data. Geostatistical functional data appear when D is a fixed subset of $\mathbb{R}^d$ with positive volume and n points $s_1, \ldots, s_n$ in D are chosen to observe the random functions $\boldsymbol{\chi}_{s_i}$, $i = 1, \ldots, n$. We say that we have a functional marked point pattern, when a complete function is observed at each point generated by a standard point process.

We focus here in the methodological issues opened around the geostatistical problems of spatial prediction and classification of functional data following Delicado *et al.* (2010) and Giraldo *et al.* (2010, 2011). In addition, and following Comas *et al.* (2011) and Mateu *et al.* (2008), we also present some issues concerning second-order characteristics in functional marked point patterns.

2 Materials and Methods

2.1 Geostatistical functional context

Let $\left\{\boldsymbol{\chi}_s(t), t \in T,\ s \in D \subset \mathbb{R}^d\right\}$ be a random function defined on some compact set T of $\mathbb{R}$. Assume that we observe a sample of curves $\boldsymbol{\chi}_{s_i}(t)$, for $t \in T$ and $s_i \in D, i = 1, \cdots, n$. It is usually assumed that these curves belong to a separable Hilbert space $\boldsymbol{H}$ of square integrable functions defined on T. We assume for each $t \in T$ that we have a second-order stationary and isotropic random process, that is, the mean and variance functions are constant and the covariance depends only on the distance among sampling sites. Formally, we assume that:

- $E(\boldsymbol{\chi}_s(t)) = m(t)$, for all $t \in T, s \in D$.
- $\mathrm{Cov}(\boldsymbol{\chi}_{s_i}(t), \boldsymbol{\chi}_{s_j}(u)) = C(h; t, u),\ s_i, s_j \in D, t, u \in T,\ h = \|s_i - s_j\|$, the Euclidean distance. If $t = u$, $\mathrm{Cov}(\boldsymbol{\chi}_{s_i}(t), \boldsymbol{\chi}_{s_j}(t)) = C(h; t)$.
- $\frac{1}{2}\mathrm{V}(\boldsymbol{\chi}_{s_i}(t) - \boldsymbol{\chi}_{s_j}(u)) = \gamma(h; t, u),\ s_i, s_j \in D, t, u \in T,\ h = \|s_i - s_j\|$. If $t = u$, $\frac{1}{2}\mathrm{V}(\boldsymbol{\chi}_{s_i}(t) - \boldsymbol{\chi}_{s_j}(t)) = \gamma(h; t)$.

The function $\gamma(h; t)$, as a function of h, is called the variogram of $\boldsymbol{\chi}(t)$.

We can use a family of point-wise linear predictors for $\boldsymbol{\chi}_{s_0}(t)$, $t \in T$, given by

$$\hat{\boldsymbol{\chi}}_{s_0}(t) = \sum_{i=1}^{n} \lambda_i(t) \boldsymbol{\chi}_{s_i}(t), \quad \lambda_1(t), \ldots, \lambda_n(t) : T \to \mathbb{R}, \tag{1}$$

For each $t \in T$, the predictor (1) has the same expression as an ordinary kriging predictor. This predictor is called the point-wise linear predictor for functional data. This modeling approach is consistent with the functional linear concurrent model (FLCM) as mentioned in Ramsay and Silverman (2005) in which the influence of each covariate on the response is *simultaneous* or *point-wise*. In our context, the covariates are the observed curves at n sites of a region and the functional response

is an unobserved function on an unsampled location. Consequently, the objective function is

$$E\|\hat{\boldsymbol{\chi}}_{s_0}(t) - \boldsymbol{\chi}_{s_0}(t)\|^2 = \int_T E\left(\hat{\boldsymbol{\chi}}_{s_0}(t) - \boldsymbol{\chi}_{s_0}(t)\right)^2 dt.$$

The predictor (1) is unbiased if $E(\hat{\boldsymbol{\chi}}_{s_0}(t)) = m(t)$, for all $t \in T$, that is, if $\sum_{i=1}^{n} \lambda_i(t) = 1$ for all $t \in T$. In this case $E\left(\hat{\boldsymbol{\chi}}_{s_0}(t) - \boldsymbol{\chi}_{s_0}(t)\right)^2 = V\left(\hat{\boldsymbol{\chi}}_{s_0}(t) - \boldsymbol{\chi}_{s_0}(t)\right)$.

We then present an approach for spatial prediction based on the functional linear point-wise model adapted to the case of spatially correlated curves. First, a smoothing process is applied to the curves by expanding the curves and the functional parameters in terms of a set of basis functions. The number of basis functions is chosen by cross-validation. Then, the spatial prediction of a curve is obtained as a point-wise linear combination of the smoothed data. The prediction problem is solved by estimating a linear model of coregionalization to set the spatial dependence among the fitted coefficients. We extend an optimization criterion used in multivariable geostatistics to the functional context. We also extend cokriging analysis and multivariable spatial prediction to the case where the observations at each sampling location consist of samples of random functions, that is, we extend two classical multivariable geostatistical methods to the functional context. Our cokriging method predicts one variable at a time as in a classical multivariable sense, but considering as auxiliary information curves instead of vectors. We also propose an extension of multivariable kriging to the functional context by defining a predictor of a whole curve based on samples of curves located at a neighborhood of the prediction site. In both cases a non-parametric approach based on basis function expansion is used to estimate the parameters, and we prove that both proposals coincide when using such an approach.

Finally, noting that classification problems of functional data arise naturally in many applications, we present methods to detect groups when the functional data are spatially correlated. Our methodology allows to find spatially homogeneous groups of sites when the observations at each sampling location consist of samples of random functions. In univariable and multivariable geostatistics various methods of incorporating spatial information into the clustering analysis have been considered. Here we extend these methods to the functional context in order to fulfill the task of clustering spatially correlated curves. In our approach we initially use basis functions to smooth the observed data, and then we weight the dissimilarity matrix among curves by either the trace-variogram or the multivariable variogram calculated with the coefficients of the basis functions.

2.2 Point pattern functional context

Despite of the relatively long history of point process theory few approaches have been performed to analyse spatial point patterns where the features of interest are functions (i.e. curves) instead of qualitative or quantitative variables. Examples of point patterns with associated functional data include forest patterns where for

each tree we have a growth function, curves representing the incidence of an epidemic over a period of time, and the evolution of distinct economic parameters such as unemployment and price rates all for distinct spatial locations. The study of such configurations permits to analyse the effects of the spatial structure on individual functions. For instance, the analysis of point patterns where the associated curves depend on time may permit the study of space-time interdependencies of such dynamic processes. However, note that time has not necessarily to be the dependent argument. Here point patterns with associated curves will be called functional marked point patterns.

Following Comas *et al.* (2011), we formulate and illustrate a new second order characteristic to analyse functional marked point patterns, the functional mark correlation function. This new statistic is a counterpart version of the mark correlation function where instead of a test function relating a quantitative mark we consider a test function involving two whole functions. This permits to analyse the spatial dependence in the functional marks. An additional mark configuration is considered by defining local characteristics in terms of LISA functions (Mateu *et al.*, 2008), and we exploit these functions to obtain functional information of the point pattern.

References

Comas C., Delicado P., Mateu J. (2011) A second order approach to analyse spatial point patterns with functional marks, *Test*, DOI: 10.1007/s11749-010-0215-1.

Delicado P., Giraldo R., Comas C., Mateu J. (2010) Statistics for spatial functional data: some recent contributions, *Environmetrics*, 21, 224-239.

Giraldo R., Delicado P., Mateu J. (2010) Continuous time-varying kriging for spatial prediction of functional data: An environmental application, *Journal of Agricultural, Biological, and Environmental Statistics*, 15, 66-82.

Giraldo R., Delicado P., Mateu J. (2011) Ordinary kriging for function-valued spatial data, *Environmental and Ecological Statistics*, DOI: 10.1007/s10651-010-0143-y.

Mateu J., Lorenzo G., Porcu E. (2008) Detecting features in spatial point processes with clutter via local indicators of spatial association, *Journal of Computational and Graphical Statistics*, 16, 968-990.

Ramsay J., Dalzell C. (1991) Some tools for functional data analysis, *Journal of the Royal Statistical Society, Series B*, 53, 539-572.

Ramsay J., Silverman B. (2005) *Functional data analysis*, Second edition, New York: Springer.

Application of a modeling system aimed at studying the impact on air quality of a waste storage fire

Roberto Giua[1], Angela Morabito[1], Annalisa Tanzarella[1]
[1] ARPA Puglia, r.giua@arpa.puglia.it

Abstract: Outdoor fires, including wildfires, prescribed burns, slash burns, and agricultural field burning can emit significant amounts of particulate matter and gaseous pollutants into the atmosphere, which can have severe effect on local and regional air quality. The aim of this study was to evaluate the impact on air quality of a waste storage fire, using a dispersion modeling system developed by Arianet©. The system includes a micrometeorological processor to reproduce the meteorology and the atmospheric turbulence and a lagrangian dispersion model to simulate the dispersion of inert pollutants on the local scale. A crucial issue was the accuracy in characterizing the emission source. Model results were compared with measured concentrations at air quality monitoring stations.

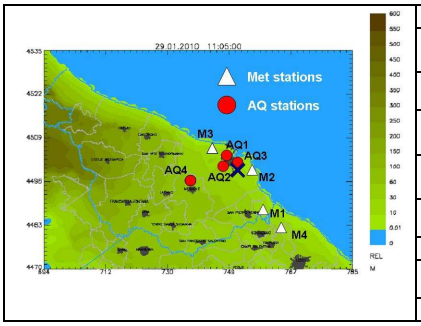
Keywords: fires, air quality, modeling system

1. Introduction

Modelling is extensively used in air quality forecasts with the aim of providing next day and near real time information to the public and for the implementation of short term action plans; it represents a useful operative support in accidental events such as wildfires, slash burns and agricultural field burning, supplying a fast view of impacts over the territory. Pollutant concentrations at ground level can be directly influenced by these events, going to affect local air quality (Yinga Q., 2009) .

The fire event analysed in this study broke out over an area near Brindisi, shown in Figure 1, in the south-eastern part of Apulia region, lapped to the east by the Adriatic sea. It started on July 9th at 14:30 pm and burned out until July 10th at 6:00 am. This period was characterized by high pressure and stable conditions, with a complete rotation of winds from northern quadrant, due to breeze circulation, during the first day. The fire developed in a waste storage, on an area of 4000 m^2 which contained mainly undefined plastic material for a total amount of about 2000 ton. The emission was considered as a point source of particulate matter (PM10), uniformly emitting during the whole period of 16 hours. Hourly meteorological data were provided by three stations (M1, M2 and M4 in Figure 1), while sounding meteorological data were provided every 12 hours by the Air Force by a station a few kilometers from Brindisi (station M3). The meteorological parameters measured by these stations are indicated in the table at the side of Figure 1.

In the area of study, PM10 is measured continuously with different sampling times, at 4 air quality monitoring sites, handled by Regional Environmental Protection Agency (ARPA). The position of each station is shown in Figure 1.



Stations	Measured variables
M1 – Torchiarolo (ground level station, ARPA)	wv (m/s), wd (deg), T (°C),
M2 – Enel (at 10m and 50m)	wv (m/s), wd (deg), T (°C),
M3 – Brindisi (radiosonde, up to 2000m)	wv (m/s), wd (deg), T (°C),
M4 – S.M. Cerrate (ground level station, ARPA)	wv (m/s), wd (deg), T (°C),
AQ1 – via Taranto (BR)	CO, C6H6, PM10 (daily), NO2, O3, SO2
AQ2 – via dei Mille (BR)	PM10, NO2, SO2
AQ3 – Sisri (BR)	CO, C6H6, PM10, NO2, SO2
AQ4 - Mesagne	PM10 (daily), NO2, SO2

Figure 1: Domain (left); study area (right) where triangles indicate meteorological stations (M1, M2, M3), black points indicate air quality stations (AQ1/AQ4), and the cross represents fire location.

2. Materials and Methods

The fire from waste disposal was simulated as a ground level point source emitting fine particulate matter (PM10). Plume rise was calculated by modified Briggs' equation, which takes into account for effective diameter of the pool fire (Fisher B., 2001). Emission was estimated considering waste disposal as a municipal refuse with a PM10 emission factor of 8 kg/tonn (EPA user guide). Disposal was considered mainly composed of plastic material with a calorific power of 31425 KJ/kg.

The impact on air quality of the fire was simulated using a dispersion modeling system developed by Arianet©. The system includes a micrometeorological processor SURFPRO (Arianet, 2007) to reproduce the atmospheric turbulence and the lagrangian dispersion model SPRAY (Arianet©, 2007) to simulate the dispersion of primary pollutants on a local scale. Modeling system can run in two modalities: the forecast run uses forecast at +24 and +48 hours, elaborated by the coupled meteorological models NCEP-RAMS (Pielke et al., 1992), the analysis run builds meteorological fields by available meteorological measures using the diagnostic wind field model MINERVE (Geai, 1987).

In order to evaluate the performance of modeling system to reproduce the evolution of pollutant over the area, the system was run in the two modalities, for 24 hours (from July 9th at 13:00 pm to 10th at 13:00 pm). In Table 1 we summarized the main features of the modeling system.

	NCEP-RAMS	Minerve	SurfPro	SPRAY
Grid		92 x 66 x 15 Δx=Δy=1 km Top domain = 6000m	92 x 66 x 15 Δx=Δy=1 km Top domain = 6000m	92 x 66 x 3 Δx=Δy=1 km Top domain = 5000m
case 1 Analysis run		Reconstruction of gridded meteorological fields by the stations **M1**, **M2** and **M3**	2D gridded turbulence fields	Transport and dispersion of pollutant
case 2 Forecast run	3D Gridded meteorological fields	Interpolation of prognostic wind fields	2D gridded turbulence fields	Transport and dispersion of pollutant

Table 1: Model configuration

3. Results

To verify the capability of the models to reproduce meteorological features of the area, a comparison between meteorological model results and measurements was performed. Figure 2 shows the evolution of modelled (forecast and analysis) and observed hourly average wind speed and direction at M1, M2 and M4 stations, for the 24 hours simulation.

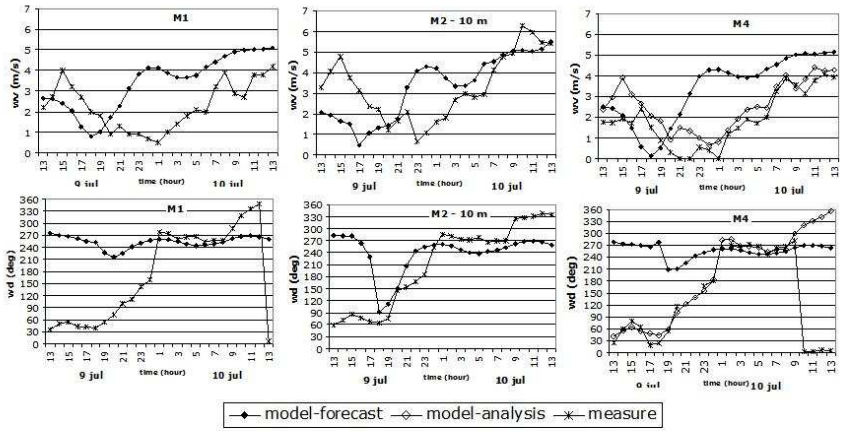


Figure 2: Evolution of modelled (forecast an analysis) and observed hourly average wind speed (top) and direction (down) at the three stations M1, M2 and M4, for the 24 hours of simulation

M1 and M2 stations were only compared with forecast simulation (case 2), while M4, being independent by the analysis run (case 1), both with forecast and analysis run. The model seemed to overestimate wind speed measured at all the stations in case 2, while a good accordance with analysis model could be appreciated at station A4. Simulated wind direction in case 2 delayed at 19:00 the rotation of the wind, while measured data at 14:00 already detected the onset of a breeze circulation. The analysis modality (case 1) reproduced quite good this rotation at station M4. In Figure 3, comparison between measured and modelled PM10 concentrations for both case 1 and 2 at AQ1, AQ2, AQ3 and AQ4 air quality stations is shown, together with - on the right of the same figure - maps of modelled average and maximum PM10 concentrations. Measurements at AQ1, AQ2 and AQ3 stations, near the city centre and the burned area, did not appear to be affected by the fire event; only AQ4 daily measurements (red line in Fig. 3) showed an increase on July 9th.

It can be observed that, in general, modelled contribution of the fire event to total PM10 concentration was quite low. Case 1 reproduces maximum concentrations on July 10th around 6:00 a.m., but it shows a peak at AQ4 station on 9th at 17:00 p.m. Case 2 shows maximum values for all the stations on 9th , between 18:00 and 20:00 p.m. A shift in time is evident between simulated PM10 in analysis (case 1) and forecast modalities (case 2), due to the reproduction of local circulation. As a consequence, modelled impact area in case 1 is substantially different from case 2. In case 1, model indicates that most significant effects were not observed near Brindisi urban area, but in S-SO direction from the ignition point, while in case 2 PM10 remain localized around the burned area.

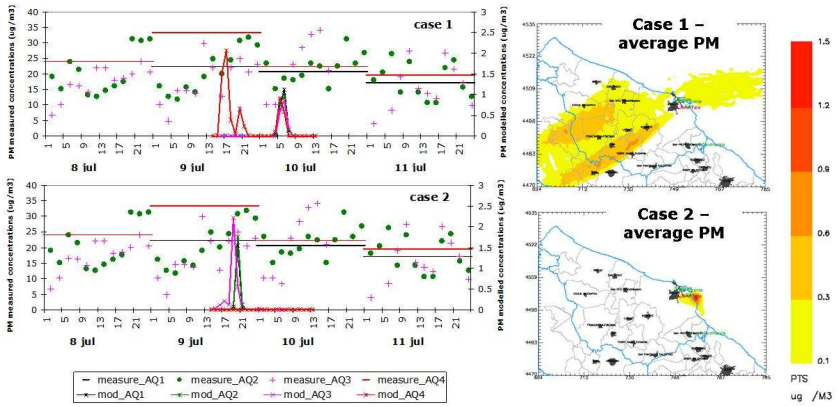


Figure 3: (left) bihourly (AQ2 and AQ3) and daily (AQ1 and AQ4) PM10 measures ($\mu g/m^3$) versus modelled (case 1 and case 2) hourly concentration; (right) average simulated ground level concentration fields of PM10 (μg/m3) for the two case studies

4. Concluding remarks

Air quality system developed by Arianet© was used to simulate the transport and dispersion of particulate matter produced by a fire event inside a waste storage. Modelled results were compared with measured data.

Results pointed out that: i) analysis modality seems to be more reliable in reproducing the evolution of pollutants over the domain, ii) the evaluation in forecast mode is rather affected by the quality of meteorological forecasting in the short term; iii) the fallout of PM10 due to the fire is low with respect to measured concentrations.

We have to stress the importance of having correct emission data, the knowledge of the type of the burned material from which to calculate the calorific power, and a good representation of meteorology able to reproduce and describe the characteristics of circulation on a local scale.

References

Arianet, 2007: SPRAY 3.1 General Description and User's Guide, R2007.08

Arianet, 2007: SURFPRO (SURrface-atmosphere interFace PROcessor) User's guide, Version 2.2.10.

EPA User Guide, AP-42 Solid Waste Disposal.

Fisher B., (2001), Modelling plume rise and dispersion from pool fires, *Atmospheric Environment*, Volume 35, 2101 – 2110.

Geai P. (1987), Methode d'interpolation and reconstitution tridimensionelle d'un champ de vent: le code d'analyse objective MINERVE, EDF/DER report HE-34/87.03.

Pielke R.A., Cotton W.R., Walko R.L., Tremback C.J., Lyons W.A., Grasso L.D., Nicholls M.E., Moran M.D., Wesley D.A., Lee T.J. and Copeland J.H. (1992), A comprehensive meteorological modelling system – RAMS. *Meteorol. Atmos. Phys.* 49, 69-91.

Yinga Q., Kleeman M. (2009) Regional contributions to airborne particulate matter in central California during a severe pollution episode, *Atmospheric Environment*, Volume 43, Issue 6, 1218-1228.

Estimation of the areas of air quality limit value exceedances on national and local scales. A geostatistical approach.[1]

Laure Malherbe[1], Maxime Beauchamp[1], Laurent Létinois[1], Anthony Ung[1]
1 : Institut National de l'Environnement Industriel et des Risques (INERIS), Direction des Risques Chroniques, Parc Technologique ALATA, 60550 Verneuil-en-Halatte, France, laure.malherbe@ineris.fr

Chantal de Fouquet[2]
2: Mines ParisTech, Centre de Géosciences, Equipe géostatistique, 35 rue Saint-Honoré, 77305 Fontainebleau, France

Abstract:
Each year Member States have to report to the European Commission on the exceedances of air quality limit values which occurred on their territory. Quantitative information is required about the areas and population exposed to such exceedances.
A probabilistic methodology for defining exceedance zones has been developed, based on preliminary air quality mapping. Atmospheric concentration fields estimated by kriging and the corresponding kriging variance are used to identify areas where the exceedance or non-exceedance can be considered as certain and areas where the situation with respect to the limit value is indeterminate. The methodology is applied on national and urban scales focusing on exceedances of PM_{10} daily limit value and NO_2 annual limit value. Results are discussed from operational perspectives.

Keywords: threshold exceedance, geostatistics, kriging, NO_2, PM_{10}

1. Introduction

In addition to reporting air quality measurement data above limit values, Member States have to provide estimates of the surface and population exposed to the observed exceedances. This study aims at developing a methodology that can be easily implemented both at national level for an overall evaluation of exceedance areas, and at local level for a more detailed assessment.

A two-stage methodology is proposed. It first involves estimating concentrations over the domain of interest and computing the estimation variance. A kriging based mapping approach can be used at that stage. Non-exceedance and exceedance zones are then determined from kriging results, considering the risk of misclassifying a point.

[1] This work was funded by the French Ministry in charge of the Ecology and Sustainable Development.

The calculation steps are described in section 2. Section 3 provides application examples on national (PM_{10}) and urban (NO_2) scales. Improvement issues are discussed in the concluding part.

2. Materials and Methods

Let LV designate the considered limit value. LV = 40 µg/m^3 for NO_2 or PM_{10} annual mean concentrations; LV = 50 µg/m^3, not to be exceeded more than 35 times per year, for PM_{10} daily mean concentrations (Directive 2008/50/EC).

Let $Z(x)$ denote the concentration at location x that has to be compared to LV, $Z^*(x)$ its estimate from kriging and $\sigma_K(x)$ the kriging standard deviation. Let us take the estimation error $\varepsilon(x)$ into account, conventionally assumed to be a Gaussian process with zero mean and a standard deviation equal to $\sigma_K(x)$:

$$\varepsilon(x) = Z(x) - Z^*(x) = \sigma_K(x).T \qquad \text{with } T \rightsquigarrow \mathcal{N}(0,1) \qquad (1)$$

Evaluating whether Z(x) exceeds the limit value can be written as follows:

$$Z(x) > LV \Leftrightarrow Z^*(x) + \sigma_K(x) \cdot T > LV \Leftrightarrow T > \frac{LV - Z^*(x)}{\sigma_K(x)} \qquad (2)$$

In the proposed method, non-exceedance and exceedance areas are delimited from inequality (2), considering a non-detection probability threshold α, which is the risk of x belonging to a non-exceedance zone whereas Z(x) is above the limit value, and a false detection probability threshold β, which is the risk of x belonging to an exceedance zone whereas Z(x) is below the limit value.
If the priority is to keep the number of exceedance points wrongly included in the non-exceedance area as small as possible, then α should be set to a low value whereas a higher value may be allowed for β. Cori (2005) suggests that α be given the classical value of 5% while β could empirically be set to 1/3 to have a moderate risk of false detection. This leads to the following definitions:

- non-exceedance zone: $\{x\}$ such as $P[Z(x) > LV] < \alpha$

$$\Leftrightarrow P\left[T > \frac{LV - Z^*(x)}{\sigma_K(x)}\right] < \alpha \Leftrightarrow Z^*(x) < LV - q_{1-\alpha} * \sigma_K(x)$$
$$\Leftrightarrow Z^*(x) < LV - 1.65 * \sigma_K(x) \quad \text{for } \alpha=5\% \qquad (3)$$

- exceedance zone: $\{x\}$ such as $P[Z(x) \leq LV] < \beta$

$$\Leftrightarrow P\left[T \leq \frac{LV - Z^*(x)}{\sigma_K(x)}\right] < \beta \Leftrightarrow Z^*(x) > LV - q_\beta * \sigma_K(x)$$
$$\Leftrightarrow Z^*(x) > LV + 0.41 * \sigma_K(x) \quad \text{for } \beta=34\% \qquad (4)$$

$q_{1-\alpha}$ and q_β are the $(1-\alpha)$ and β-quantiles from the standard normal distribution.

The locations satisfying none of those conditions make the “uncertainty zone”. In section 3 this formal approach is compared to a more empirical methodology previously developed for identifying exceedances of PM_{10} daily limit value and rapidly answering to urgent regulatory requests (Malherbe and Cárdenas, 2009; GT Zones sensible, 2010).

3. Results

National level. On French scale, daily PM_{10} and annual NO_2 concentrations are estimated on a 1 km x 1km grid by combining surface observations from background monitoring stations with outputs from the chemistry-transport model CHIMERE (resolution : about 10 km). For NO_2, which is mainly related to local emission sources, additional high resolution variables, precisely NO_x emission density and population density within a 2-km radius, are introduced in the kriging as external drift.

Figure 1 shows the example of one polluted day during a PM_{10} event (April 2009). Results are provided both for the methodology described in section 2 and the more empirical methodology in which only two states are defined (hypothesis (1) being unchanged):

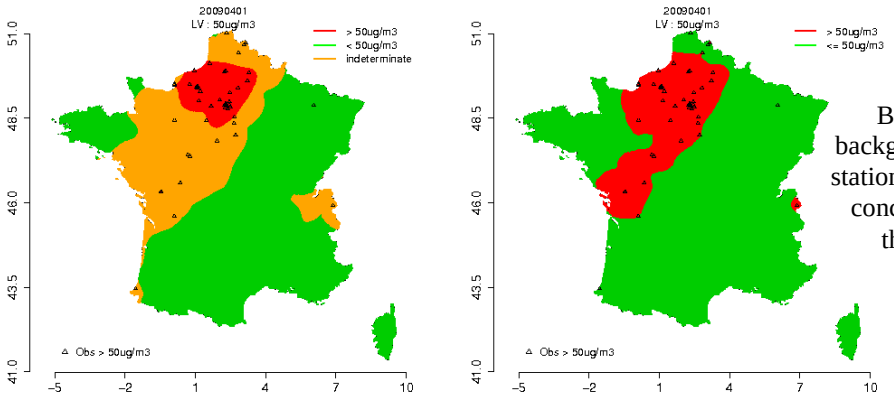
- exceedance: $\{x\}$ such as $P[Z(x) \leq LV] < \eta$ with η: false detection probability threshold

$$\Leftrightarrow P\left[T \leq \frac{LV - Z^*(x)}{\sigma_K(x)}\right] < \eta \Leftrightarrow Z^*(x) > LV - q_\eta * \sigma_K(x) \qquad (5)$$

q_η is empirically adjusted by comparing the annual numbers of exceedances estimated by cross-validation at the monitoring sites with the actual observed numbers. In this application q_η has been set to approximately 0.52, which amounts to defining a false-detection probability threshold of 70% and taking a cut-off value lower than LV.

- non-exceedance: $\{x\}$ making the complementary set, i.e. :

$$\{x\} \text{ such as } Z^*(x) \leq LV - q_\eta * \sigma_K(x) \qquad (6)$$



Black triangles: background monitoring stations with daily mean concentration higher than 50 µg/m^3.

Figure 1: PM_{10}. Exceedance of the daily threshold (50 µg/m^3) on a highly polluted day. Left: newly formalized methodology. Right: empirical methodology.

Local level

Exceedance areas defined for NO_2 on national scale are very small since NO_2 exceedances mostly occur at traffic-related sites. On local scale, detailed concentration maps accounting for both background and roadside pollution can be established from passive sampling surveys, using high resolution auxiliary variables and additional information about traffic emissions and distance to the roads (Malherbe et al., 2008).
Results obtained for the French city of Montpellier are displayed in Figure 2. During year 2007 an extensive sampling campaign was carried out in this city by the local agency Air Languedoc-Roussillon. The sampling dataset includes eight 14-day periods of measurement at background and traffic sites.

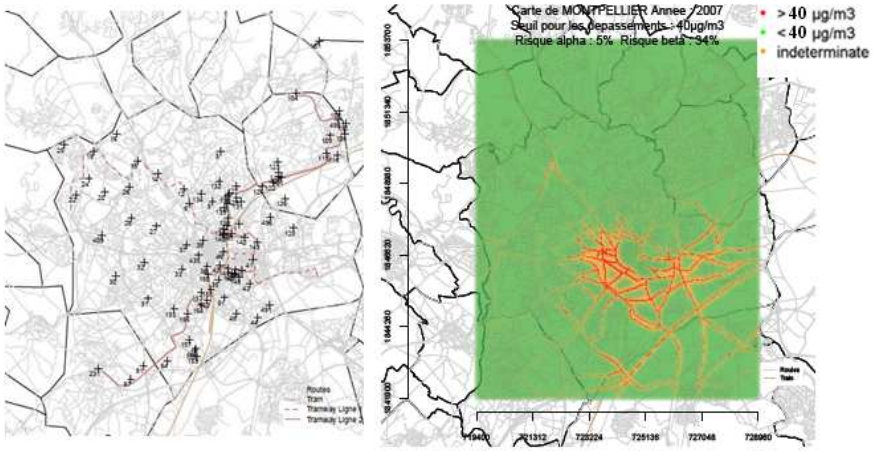


Figure 2: NO_2. Left: sampling points in Montpellier - year 2007 (source of the data: Air Languedoc Roussillon). Right: exceedance of the annual limit value (40 µg/m^3).

4. Concluding remarks

Annual reporting to the European Commission but also the working out of local air quality plans require the delimitation of areas where atmospheric concentrations do not comply with environmental objectives. A first approach was developed with a view to rapidly producing realistic exceedance maps for PM_{10}. The identified areas are consistent with observed exceedances but might somewhat be overestimated, especially where or when kriging variance is high. The notion of exceedance and non-exceedance was then formalized making some conventional assumptions due to operational constraints. The advantage of this second approach is that it distinguishes the non-detection and false detection probability thresholds which can be adjusted according to the objectives of the study. However, a remaining issue is the way of addressing the uncertainty area. In the end authorities and decision makers will rather have a single figure (spatial extent of the exceedance) than an interval of values. Among envisaged solutions are the refining of the uncertainty area and its inclusion in the exceedance area.

References

Cori A. (2005). Définition de zones homogènes vis-à-vis du dépassement de seuil pour la concentration en ozone. Mémoire de stage de l'Ecole des Mines de Paris effectué à Air Normand, troisième partie.

Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. http://eur-lex.europa.eu

Groupe de travail national *Zones sensibles* (2011). Méthodologie de définition des zones sensibles. http://www.lcsqa.org

Malherbe L., Cárdenas G., (2009). Evaluation des zones touchées par les dépassements de valeurs limites. Note méthodologique. Version 1. http://www.lcsqa.org

Malherbe L., Cárdenas G., Colin P., Yahyaoui A. (2008). Using different spatial scale measurements in a geostatistically based approach for mapping atmospheric nitrogen dioxide concentrations, *Environmetrics*, 19, 751–764

Modeling pollutant threshold exceedance probabilities in the presence of exogenous variables[1]

Rosaria Ignaccolo
Università di Torino, Italy, ignaccolo@econ.unito.it

Dana Sylvan
Hunter College of the City University of New York, USA

Michela Cameletti
Università di Bergamo, Italy

Abstract: Many studies link exposure to various air pollutants to respiratory illness, making it important to identify regions where such exposure risks are high. One way of addressing this problem is by modeling probabilities of exceeding specific pollution thresholds. In this paper, we consider particulate matter with diameter less than 10 microns (PM_{10}) in the North-Italian region Piemonte. The problem of interest is to predict the daily exceedance of 50 micrograms per cubic meter of PM_{10} based on air pollution data, geographic information, as well as exogenous variables. We use a two-step procedure involving nonparametric modeling in the time domain, followed by spatial interpolation. Resampling schemes are employed to evaluate the uncertainty in these predictions.

Keywords: exceedance probability map, air pollution, space-time modeling

1 Motivation, background, data

It is well known that high levels of pollutants in the ambient air have adverse effects on human and environmental health. Environmental directives have been issued in order to account for such potential dangers, setting limit values for various air pollutants. By estimating the probability to exceed a fixed value of a given pollutant, we can identify areas where the risk to exceed such limit values is high. Past environmental studies focused on mean behavior revealed that inclusion of exogenous variables may lead to better estimators and predictors of pollutant concentrations. It seems therefore natural to expect that including additional information, such as meteorological and orographical variables might improve daily predictions of exceedance probabilities. In this study we extend the methodology introduced

[1]Ignaccolo's work was partially supported by Regione Piemonte, while Sylvan's research was funded in part by the PSC-CUNY Research Award 63147-00-41.

in Draghicescu and Ignaccolo (2009) by including exogenous variables. Our case study considers daily PM_{10} concentrations (in $\mu g/m^3$) measured from October 2005 to March 2006 by the monitoring network of Piemonte region (Italy) containing 24 sites. As covariates we use daily maximum mixing height (HMIX, in m), daily mean wind speed (WS, in m/s), daily emission rates of primary aerosols (EMI, in g/s), altitude (A, in m) and coordinates (UTMX and UTMY, in km). Note that the time-varying variables are obtained from a nested system of deterministic computer-based models implemented by the environmental agency ARPA Piemonte. For a complete description and preliminary analysis of the data we refer to Cameletti et al. (2010).

2 Theoretical Framework

Let $D \subset \mathbf{R}^2$, and assume that at each location $s \in D$ we observe a temporal process $X_s(t) = G_s(t, Z_s(t))$, where G_s is an unknown transformation, Z_s is a standardized stationary Gaussian process with $\gamma_s(l) := cov(Z_s(t), Z_s(t+l))$, such that $\sum_{l=-\infty}^{\infty} |\gamma_s(l)| < \infty$. For fixed $x_0 \in \mathbf{R}$, define the exceedance probability

$$\mathbf{P}_{x_0}(t, s) = P(X_s(t) \geq x_0). \tag{1}$$

Clearly $\mathbf{P}_{x_0}(t, s)$ takes values in $[0, 1]$ and is non-increasing in x_0. The problem of interest is to predict $\mathbf{P}_{x_0}(t, s^*)$ at location $s^* \in D$ where there are no observations and at any time t, based on observations of the process $X_s(t)$ at n time points and m spatial locations.

In the *first step* we use the methodology proposed in Draghicescu and Ignaccolo (2009). For each site s, we model the temporal risks non-parametrically, by using the Nadaraya-Watson kernel estimator

$$\hat{\mathbf{P}}_{x_0}(t, s) = \frac{\sum_{i=1}^{n} K\left(\frac{t_i - t}{b_t}\right) 1_{\{X_s(t_i) \geq x_0\}}}{\sum_{i=1}^{n} K\left(\frac{t_i - t}{b_t}\right)}, \tag{2}$$

where K is a kernel function. The temporal bandwidth b_t should not depend on the threshold x_0, in order for the resulting estimator to be non-increasing. In what follows, the threshold x_0 is considered fixed and, to keep notation simple, we write b instead of b_t. In the *second step*, we use universal kriging with exogenous variables to predict the exceedance probability field at a location $s^* \in D$ where there are no observations. Since linear interpolation does not guarantee that the resulting exceedance probability estimator takes values in the interval $[0, 1]$, we first apply a $1:1$ transformation and consider $\hat{Q}_{x_0}(t, s) = \Phi^{-1}(\hat{\mathbf{P}}_{x_0}(t, s))$ which is defined on $\mathbf{R}$, where $\Phi(\cdot)$ is the standard Normal cumulative distribution function. After performing kriging on the transformed field $\hat{Q}_{x_0}(t, s)$, we obtain the desired exceedance probability maps by inversion: $\hat{\mathbf{P}}_{x_0}(t, s) = \Phi(\hat{Q}_{x_0}(t, s))$. For fixed time point t and location s_i, we consider the model

$$\hat{Q}_{x_0}(t, s_i) = \beta E(t, s_i) + w(t, s_i), \tag{3}$$

where $E(t, s_i)$ is a vector of exogenous variables at time t and location s_i, β is the vector of "slopes", and $w(t, s)$ is a zero-mean second-order stationary spatial process for any $s \in D \subset \mathbf{R}^2$. Time point t is fixed, and the spatial covariance is denoted by $C(t, ||s_i - s_j||) := Cov\big(w(t, s_i), w(t, s_j)\big)$. We then use the Matèrn class to model this covariance function: $C(t, ||s_i - s_j||) = \frac{\sigma_t}{2^{\nu_t - 1}\Gamma(\nu_t)} \Big(\frac{2\sqrt{\nu_t}||s_i - s_j||}{\rho_t}\Big)^{\nu_t} \mathcal{K}_{\nu_t}\Big(\frac{2\sqrt{\nu_t}||s_i - s_j||}{\rho_t}\Big)$. The parameter $\nu_t > 0$ characterizes the smoothness of the process, σ_t denotes the variance, and ρ_t measures how quickly the correlation decays with distance. For each t, the parameters of the Matèrn covariance are estimated by weighted least squares. The best linear unbiased predictor of the transformed field at location $s_0 \in D$ is obtained via universal kriging (Gaetan and Guyon 2010, p. 44) as

$$\hat{Q}^*_{x_0}(t, s_0) = \hat{\beta} E(t, s_0) + w^*(t, s_0). \tag{4}$$

Here $\hat{\beta}$ is the generalized least squares estimate of the trend coefficients and $w^*(t, s_0) = \sum_{i=1}^{m} \lambda_i \hat{w}(t, s_i)$ is the simple kriging predictor, with $\hat{w}(t, s_i) = \hat{Q}_{x_0}(t, s_i) - \hat{\beta} E(t, s_i)$. The weights λ_i, $1 \leq i \leq m$ are completely determined by the covariance function parameters ν_t, ρ_t, and σ_t. The standard error of $\hat{Q}^*_{x_0}(t, s_0)$ can be also expressed in terms of the interpolation parameters λ_i. However, this standard error may not be completely accurate since the Matèrn parameters are estimated from the same data thus adding uncertainty, and the error induced by the first step of our procedure is not considered. For these reasons, we use block bootstrap (Buhlmann, 2002) to take into account all the uncertainty sources.

3 Results

In this case study on the North Italian region Piemonte we used data at $m = 24$ sites and $n = 182$ days. The PM_{10} threshold was set to $x_0 = 50\ \mu g/m^3$. The computations were done in `R`, using the `gstat` package (Pebesma, 2004). Regarding the bootstrap, we sampled with replacement $k = 13$ blocks of length $l = 14$ from the $(n - l + 1)$ possible overlapping blocks. We chose $l = 14$ empirically. A temporal window of two weeks captures the meteorological and air pollution patterns well. Also, by trying other values we did not get significantly different results. In future research we plan to generalize the methodology of Li et al. (2007) to more complex dependencies. The block sampling was then repeated B times, yielding the B bootstrap samples. Bootstrap replicated exceedance probability maps were obtained by performing the first and second steps on each bootstrap sample. In the spatial interpolation step we used a 56×72 regular grid covering Piemonte. Based on the distribution of the B bootstrap replications, we obtained the quantile maps together with the standard errors of the exceedance probability predictions. In our computations we used $B = 500$ bootstrap replications. Maps of the 10th, 50th and 90th percentiles of the exceedance probability bootstrap distribution for March 5, 2006 are showed in Figure 1, identifying increased risks around the metropolitan area of Torino.

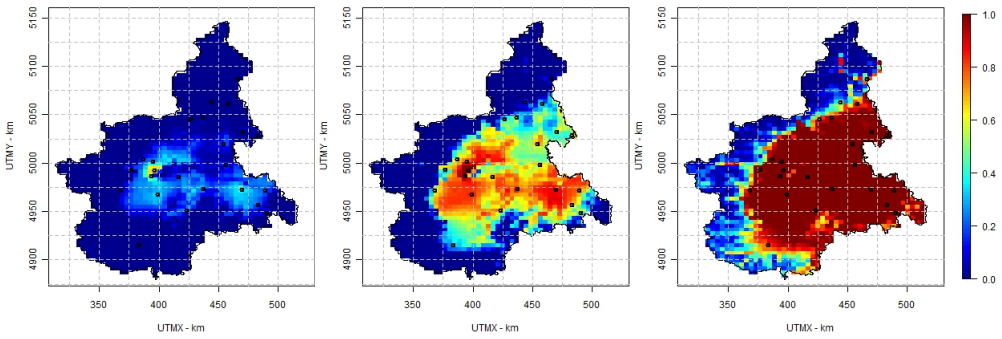


Figure 1: Maps of the bootstrap predicted 50 $\mu g/m^3$ PM_{10} exceedance probabilities on March 5th, 2006: 10th (left), 50th (center) and 90th percentile (right).

4 Discussion

This work is a continuation of Draghicescu and Ignaccolo (2009), where preliminary exceedance probability maps were obtained based on a two-step procedure. Seasonal (winter and summer) maps were quite good, however, the daily exceedance probability maps did not seem to reflect the true air pollution spatial patterns well. By introducing exogenous variables we were now able to obtain more reasonable spatial patterns for air pollution risks in Piemonte. In addition, we obtained confidence regions by estimating uncertainty in our predictions through bootstrap. It seems though that the standard errors might be too large, possibly because the shuffling in the block bootstrap did not respect the temporal evolution of the process. Our ongoing research is focused on improving these confidence bands by considering seasonal time series bootstrap.

References

Buhlmann P. (2002). Bootstraps for Time Series, *Statistical Science*, 17, 1, 52-72.

Cameletti M., Ignaccolo R., Bande S. (2010). Comparing air quality statistical models, *GRASPA Working Papers*, 40 (downloadable at `www.graspa.org`).

Draghicescu D., Ignaccolo R. (2009). Modeling threshold exceedance probabilities of spatially correlated time series, *Electronic Journal of Statistics*, 3, 149-164.

Gaetan C., Guyon X. (2010). *Spatial Statistics and Modelling*. Springer.

Li, B., Genton, M.G., Sherman, M. (2007). Nonparametric assessment of properties of space-time covariance functions, *JASA*, 102, 736–744.

Pebesma E.J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, 30, 683–691.

Using the SPDE approach for air quality mapping in Piemonte region [1]

Michela Cameletti
Università di Bergamo, Bergamo (I), michela.cameletti@unibg.it

Finn Lindgren, Daniel Simpson and Håvard Rue
Norwegian University of Science and Technology, Trondheim (N)

Abstract: In this work we consider a geostatistical spatio-temporal model for PM_{10} concentration (particulate matter with an aerodynamic diameter of less than 10 μm) in the North-Italian region Piemonte. The model involves a Gaussian Field (GF) affected by a measurement error and a state process with a first order autoregressive dynamics and spatially correlated innovations. The main goal of this work is to propose an estimating and mapping strategy for such a model. This proposal is based on the work of Lindgren et al. (2011) that provides an explicit link between GFs and Gaussian Markov random fields (GMRF) through the Stochastic Partial Differential Equations (SPDE) approach. Thanks to the `R` library named `INLA`, the SPDE approach can be easily implemented providing results in reasonable computing time (with respect to other MCMC algorithms). For these reasons, the SPDE approach is proved to be a powerful strategy for modeling and mapping complex spatio-temporal phenomena.

Keywords: spatio-temporal model, Integrated Nested Laplace Approximation, big n problem.

1 Introduction

In the geostatistical approach, data coming from monitoring networks are assumed to be realizations of a continuously indexed spatial process changing in time $\mathcal{Y}(s,t) = \{y(s,t) : (s,t) \in \mathcal{D} \subseteq \mathbb{R}^2 \times \mathbb{R}\}$, also named *random field*. These realizations are used to make inference about the process and to predict it at desired locations (i.e. kriging). Generally, we deal with a Gaussian field (GF) that is completely specified by its mean and spatio-temporal covariance function $Cov\left(y(s,t), y(s',t')\right) = \sigma^2\mathcal{C}((s,t),(s',t'))$, defined for each (s,t) and $(s',t') \in \mathbb{R}^2 \times \mathbb{R}$. Even if the geostatistical approach is very intuitive, it suffers from the so-called "big n problem" that arises especially in case of large datasets in space and time. In particular, this computational challenge arises in the Bayesian framework where matrix operations are

[1] Cameletti's research was funded in part by Lombardy Region under "Frame Agreement 2009" (Project EN17, "Methods for the integration of different renewable energy sources and impact monitoring with satellite data").

computed iteratively for MCMC algorithms. A possible solution for facing this issue consists in representing a Matérn random field - a continuously indexed GF with a Matérn covariance function - as a discretely indexed random process, i.e. a Gaussian Markov Random Field (GMRF, Rue et al. (2005)). This proposal is based on the work of Lindgren et al. (2011), where an explicit link between GFs and GMRFs is provided through the Stochastic Partial Differential Equations (SPDE) approach. The key point is that the spatio-temporal covariance function and the dense covariance matrix of a GF are substituted, respectively, by a neighbourhood structure and by a sparse precision matrix, that together define a GMRF. The advantage of moving from a GF to a GMRF stems from the good computational properties that the latter enjoys. In fact, GMRFs are defined by a precision matrix with a sparse structure that makes it possible to use computationally effective numerically methods, especially for fast matrix factorization. Moreover, when dealing with Bayesian inference for GMRFs, it is possible to make use of the Integrated Nested Laplace Approximation (INLA) algorithm proposed by Rue et al. (2009) as an alternative to MCMC methods. The most outstanding advantage of INLA is computational because it produces almost immediately accurate approximations to posterior distributions, also in case of complex models. Thus, the joint use of the SPDE approach together with the INLA algorithm can be a powerful solution for overcoming the computational problems of spatio-temporal GFs.

2 The spatio-temporal model and the SPDE approach

Let $y(s_i, t)$ denote the PM_{10} concentration measured at station $i = 1, \ldots, d$ and day $t = 1, \ldots, T$. We assume the following measurement equation

$$y(s_i, t) = \boldsymbol{z}(s_i, t)\boldsymbol{\beta} + x(s_i, t) + \varepsilon(s_i, t) \tag{1}$$

where $\boldsymbol{z}(s_i, t) = (z_1(s_i, t), \ldots, z_p(s_i, t))$ denotes the vector of p covariates for site s_i at time t, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the coefficient vector. Moreover, $\varepsilon(s_i, t) \sim N(0, \sigma_\varepsilon^2)$ is the measurement error defined by a Gaussian white-noise process, serially and spatially uncorrelated. Finally, $x(s_i, t)$ is the so-called state process, i.e. the true unobserved level of pollution. It is supposed to be a spatio-temporal GF that changes in time with a first order autoregressive dynamics with coefficient a and coloured innovations, given by

$$x(s_i, t) = ax(s_i, t-1) + \omega(s_i, t) \tag{2}$$

where $x(s_i, 0) \sim N(0, \sigma_0^2)$ and $|a| < 1$. In particular, the zero-mean Gaussian process $\omega(s_i, t)$ is supposed to be i.i.d. over time and is characterized by the following spatio-temporal covariance function $Cov(\omega(s_i, t), \omega(s_j, t')) = \sigma_\omega^2 \mathcal{C}(h)$ for $t = t'$ and $i \neq j$. The purely spatial correlation function $\mathcal{C}(h)$ depends on the location s_i and s_j only through the Euclidean spatial distance $h = ||s_i - s_j|| \in \mathbb{R}$; thus,

the process is supposed to be second-order stationary and isotropic. The spatial correlation function $\mathcal{C}(h)$ is defined in the Matérn class and is given by $\mathcal{C}(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}} (\kappa h)^{\nu} K_{\nu} (\kappa h)$, with K_{ν} denoting the modified Bessel function of second kind and order $\nu > 0$. The parameter ν measures the degree of smoothness of the process. Instead, $\kappa > 0$ is a scale parameter whose inverse $1/\kappa$ can be interpreted as the range, i.e. the distance at which the spatial correlation becomes almost null. Collecting all the observations measured at time t in a vector denoted by $\boldsymbol{y}_t = (y(s_1, t), \ldots, y(s_d, t))'$, it follows that (1) and (2) can be written as

$$\boldsymbol{y}_t = \boldsymbol{z}_t \boldsymbol{\beta} + \boldsymbol{x}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \sigma_{\varepsilon}^2 I_d) \tag{3}$$

$$\boldsymbol{x}_t = a\boldsymbol{x}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim N(\mathbf{0}, \Sigma = \sigma_{\omega}^2 \tilde{\Sigma}) \tag{4}$$

where $\boldsymbol{z}_t = (\boldsymbol{z}(s_1, t)', \ldots, \boldsymbol{z}(s_d, t)')'$ and $\boldsymbol{x}_t = (x(s_1, t), \ldots, x(s_d, t))'$ with $\boldsymbol{x}_0 \sim N(\mathbf{0}, \sigma_0^2 I_d)$. Moreover, the d-dimensional correlation matrix $\tilde{\Sigma}$ is defined as $\tilde{\Sigma} = \mathcal{C}\left(\|s_i - s_j\|\right)_{i,j=1,\ldots,d}$, and the correlation function $\mathcal{C}(.)$ is parameterized by κ and ν.

The aim of the SPDE approach is to find a GMRF, with local neighbourhood and sparse precision matrix $\mathbf{Q}$, that best represents the Matérn field $\omega(s, t)$. As described in Lindgren et al. (2011), this results in expressing the Matérn field as a linear combination of basis functions defined on a triangulation of the domain $\mathcal{D}$ using n vertices. It follows that, for each time point t the term $\boldsymbol{\omega}_t$ introduced in Eq.(4) is represented through the GMRF $\tilde{\boldsymbol{\omega}}_t \sim N(\mathbf{0}, \mathbf{Q}_S^{-1})$, whose n-dimensional precision matrix $\mathbf{Q}_S$ comes from the SPDE representation and is computed using Eq.(10) of Lindgren et al. (2011). In particular, this defines an explicit mapping from the parameters of the GF covariance function (κ and ν) to the elements of the precision matrix $\mathbf{Q}_S$ of the GMRF.

Parameter estimation and mapping are carried out in a full Bayesian framework using the INLA algorithm which is an alternative to MCMC for getting the approximated posterior marginals for the latent variables (all over the triangulated domain) as well as for the hyperparameters (see Rue et al., 2009).

3 Data and results

In the case study on the North-Italian region Piemonte, we analyze log-transformed daily PM_{10} concentration (in $\mu g/m^3$) measured from October 2005 to March 2006 (for a total of $T = 182$ days) by $d = 24$ monitoring stations. In addition, we consider the following covariates proved to have a significative effect on pollutant dispersion: daily maximum mixing height (HMIX, in m), daily mean wind speed (WS, m/s), daily emission rates of primary aerosols (EMI, in g/s), daily mean temperature (TEMP, in K), altitude (A, in m) and coordinates (UTMX and UTMY, in km). For a complete description and preliminary analysis of the data we refer to Cameletti et al. (2010). We perform the analysis using the `R` library named `INLA` (`www.r-inla.org`) considering $n = 600$ triangle vertices and $\nu = 1$. Figure 1 displays the posterior mean of PM_{10} (on the logarithmic scale) for January 29th, 2006 together with an

uncertainty measure (standard deviation). As expected, higher levels of particulate matter pollution are detected in the metropolitan areas of the region located near the main cities (Torino, Vercelli and Novara) and moving eastwards toward Milan.

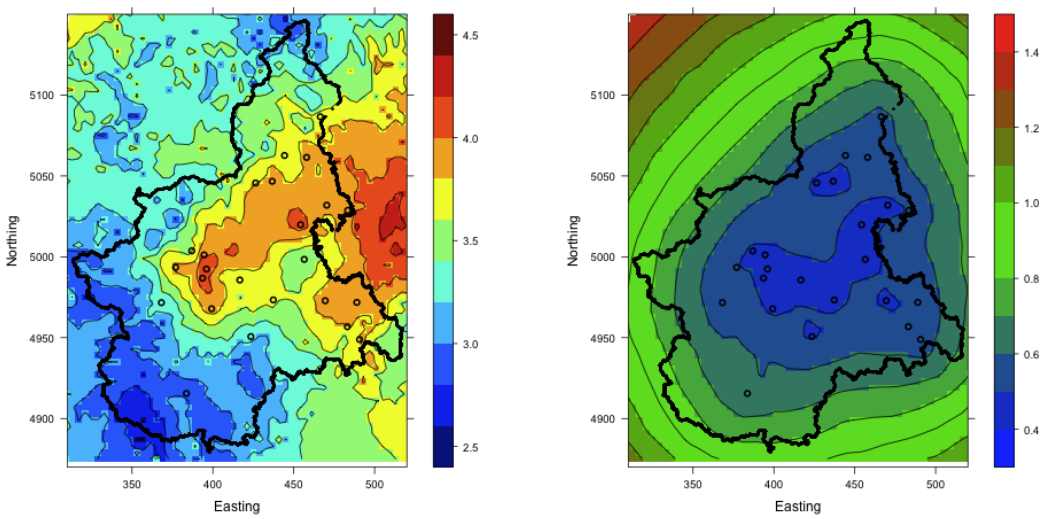


Figure 1: Map of the PM_{10} posterior mean on the logarithmic scale (left) and standard deviation (right) for January 29th, 2006.

4 Concluding remarks

In this work we present a modeling strategy - based on the SPDE approach - for a geostatistical spatio-temporal model, and show the results for a case study on air quality in Piemonte. Our ongoing research is focused on the change of support problem in order to include covariates with different spatial support in our modeling framework.

References

Cameletti M., Ignaccolo R., Bande S. (2010). Comparing air quality statistical models, *GRASPA Working Papers*, 40 (downloadable at `www.graspa.org`).

Lindgren F., Rue H., Lindström J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the SPDE approach (with discussion). *J. R. Statist. Soc. B*, 73.

Rue H., Martino S., and Chopin N. (2009) Approximate Bayesian inference for latent Gaussian model by using integrated nested Laplace approximations (with discussion). *J. R. Statist. Soc. B*, 71, 319-392.

Rue H., Held L. (2005) *Gaussian Markov Random Fields. Theory and Applications.* Chapman & Hall.

A generalization of the Incidence Function Model for metapopulations with fluctuating behaviour:an application to *Lymantria dispar* (L.) in Sardinia.

Antonella Bodini
Institute of Applied Mathematics and Information Technology
(CNR-IMATI, Milano), antonella.bodini@mi.imati.cnr.it

Gianni Gilioli
Department of Biomedical Sciences and Biotechnology,
University of Brescia

Arturo Cocco, Andrea Lentini, and
Pietro Luciano
Department of Plant Protection, University of Sassari

Abstract: Most of the analysis and modeling approaches on gypsy moth population dynamics have been applied to a continuous spatial dimension, and therefore they do not account for the possible role of highly fragmented forest stands on pest dynamics. Spatially explicit metapopulation models show some advantages in representing the spatio-temporal metapopulation dynamics in fragmented habitats. In this work, the most popular of these models has been extended to take into account periodicity in the pest dynamics. Data on the gypsy moth *Lymantria dispar* (L.) (Lepidoptera Lymantriidae), one of the main oak forest defoliators in the Holarctic Region, referring to the period 1980-2010 in Sardinia (Italy) are analyzed.

Keywords: Spatially explicit metapopulation models, Incidence Function Model, *Lymantria dispar* (L.) (Lepidoptera Lymantriidae)

1. Introduction

The gypsy moth *Lymantria dispar* (L.) (Lepidoptera Lymantriidae) population dynamics modeling has a long history. Models developed range in complexity and approaches that have been used, from statistical models (Zhou & Liebhold, 1995; Cocco *et al*., 2010) to simulation models based on complex assumptions on ecological processes (Sharov & Colbert, 1996). In many situations, especially in the oak forests of the Mediterranean basin, the host plants for the gypsy moth are not continuous. However, the role of habitat fragmentation in determining the pattern of gypsy moth population dynamics has not been carefully addressed. Analyses of spatial heterogeneity are either based on correlations that take into account details of landscapes and their effect on population processes (Hunter, 2002) or on metapopulation models that deal with the occurrence of individual populations in an ensemble of habitat fragments (Tscharntke & Brandl, 2004). Spatially explicit metapopulation models could be of great importance to pest managers for their contribution to a better understanding of

how the spatial arrangements of fields or forest stands can influence the population dynamics. Despite these promises and the fact that metapopulation models have been originally proposed for pests, they remain a widely used tool in conservation biology but receive little attention in pest control (Hunter, 2002).
In this paper, we propose a modelling approach to *L. dispar* metapopulation dynamics and apply it to a dataset of gypsy moth abundance recorded in Sardinia (Italy). Model simulations are performed and the obtained dynamics are evaluated in their capability to capture the most significant properties of spatio-temporal population dynamics patterns. The proposed model significantly improves the results obtained by Gilioli *et al.* (2011a).

2. Materials and Methods

Data. Gypsy moth population dynamics were recorded in the period 1980-2010 in the main cork, holm and pubescent oak areas of Sardinia based on 282 monitoring sites (Luciano, 1989; Cocco *et al*., 2010). Each monitoring site has been considered as the centroid of a patch, the basic environmental unit in which the local dynamics of colonization and extinction occur. Patches connected by fluxes of migrant larvae are considered belonging to the same macroarea (MA). MAs are separated by physical or ecological barriers, and fluxes among MAs can be considered negligible. Five MAs were identified: the results on MA 2 are presented here.
Model description. The Incidence Function Model (IFM; Hanski 1994) is based on presence/absence data of a species in a highly fragmented landscape. The process of occupancy of patch *i* is described by a first-order Markov chain with two states, {0, 1} (empty and occupied, respectively). Following Hanski, the colonization probability of patch *i* at time *t*, $C_i(t)$, is defined to be a sigmoidal function increasing with connectivity

$$C_i(t) = \Delta_i^2(t) / (\Delta_i^2(t) + y^2) \qquad (1)$$

where $\Delta_i(t) = \Sigma \{o_j (t) \exp (- \alpha \times r_{ij} \times d_{ij})A_j : j \neq i\}$ is the connectivity of patch *i* at time *t*, A_j is the area of patch *j*; d_{ij} is the centroid-to-centroid (Euclidean) distance between patches *i* and *j*; r_{ij} corrects the Euclidean distance by taking into account possible disturbances (presence of a different host species, grazing, etc.); *y* describes the colonization ability of the species, α is a positive constant setting the survival rate of migrants over the distance.
In this paper, the extinction probability of a population in patch *i* at time *t* is assumed to be a sigmoidal function increasing with the recent history of the patch

$$E_i (t ; K) = h_i^2(t ; K) / (h_i^2(t; K) + x(t)^2) \qquad (2)$$

where $h_i(t ; K) = \Sigma\{o_i(k) : k = t, t\text{-}1, \dots , t\text{-}K\}$, $o_i(k) = 1$ if at time *t* patch *i* is occupied and $o_i(k) = 0$ otherwise; $x(t) = \gamma + \beta \, \mathrm{Sin}^2 (\theta t + \pi \, s/\theta)$ is a sinusoidal function accounting for periodicity in local dynamics, which is common to all the patches in the same MA.
Parameters α, γ, β, θ, *s* and *y* are estimated by maximization of the pseudo-likelihood corresponding to the initial distribution given by the first observed metapopulation state (Moilanen, 1999).

3. Results

Figure 1 clearly shows the periodic behavior of the observed proportion of occupied patches (green line). Data of the period 1980-1983 have been discarded due to the high number of missing data. The first population peak is reported in 1990, the other population outbreak occurs in 1997. After 2000, the pattern of fluctuations displays less regularity, which can be partially explained by pest control treatments carried out to reduce the impact of gypsy moth infestation in the sites where outbreaks started (foci).

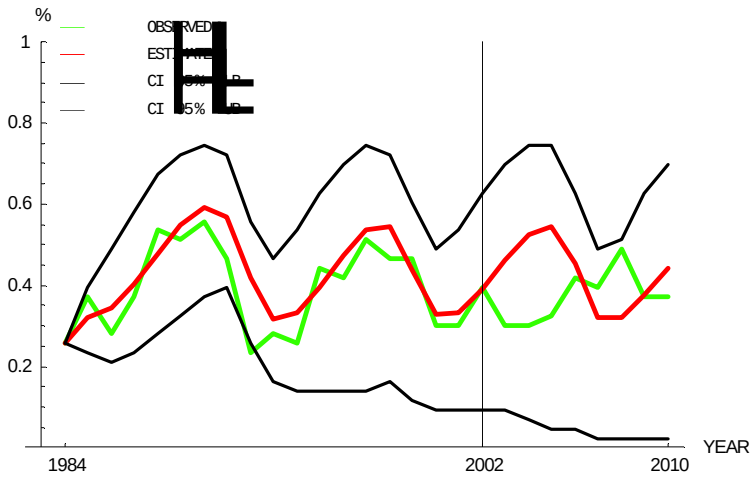


Figure 1: Observed and mean estimated proportion of occupied patches. Data of the period 1984-2002 have been used for estimation. After 2002, groups of patches received pest control treatments.

The estimated period is 7.1 years (θ = 0.44), K=5, α=0.01767, β=10.405, γ=3.031, s=0.98 and y=86.1. To compare model outputs and observed incidences, 10,000 simulations have been carried out, starting from the first year of data (1984).
Figure 1 compares the observed fraction of occupied sites and the mean estimated fraction, obtained from simulations. Confidence intervals have been obtained by computing the symmetric percentiles (0.025, 0.975) of the simulated values, for each time t. Before 2002, the observed data seem to be well represented by the model. The model behaviour after 2002 differs from observations as the populations dynamics are influenced by the pest control treatments. According to a few preliminary results, the estimated model seems to be able to adapt to the implementation of pest management strategies.

4. Concluding remarks

The major advantage offered by the metapopulation model developed here is the possibility of describing temporal trends of population dynamics in phase with

observations. In particular, the increase in the incidence at MA-level for population following a latency period, is well described by the model. This has important implications for sampling strategies as well, leading to the possibility of using a binomial sample design for management purposes, by defining the state of presence/absence of gypsy moth population abundance instead of counting egg masses. The description of variation in population incidence could allow to obtain a descriptor of the increase in the risk of population outbreaks. Different management strategies could be evaluated according to the approach proposed by Gilioli *et* al. (2008) and Gilioli *et al*. (2011b) and based on the IFM and the Kullback-Leibler divergence (Kullback and Leibler, 1951).

References

Cocco A., Cossu Q.A., Erre P., Nieddu G., Luciano P. (2010) Spatial analysis of gypsy moth populations in Sardinia using geostatistical and climate models, *Agricultural and Forest Entomology*, 12, 417-426.

Gilioli G., Bodini A., Baumgärtner J. (2008) A novel approach based on Information Theory to rank conservation strategies: an application to amphibian metapopulations, *Animal Conservation*, 11, 453-462.

Gilioli G., Bodini A., Cocco A., Lentini A., Luciano P. (2011a) Analysis and modelling of *Lymantria dispar* (L.) metapopulation dynamics in Sardinia, in *Proceedings of IOBC Working Group Integrated Protection in Oak Forest*, 6th Meeting Tempio Pausania (Italy),October 4th-10th. *In press*.

Gilioli G., Bodini A., Baumgärtner J. (2011b) Ranking control strategies for pest metapopulation management: an application to the Pine processionary moth, *submitted*.

Hanski I. (1994) A practical model of metapopulation dynamics, *Journal of Animal Ecology*, 63, 151-162.

Hunter M.D. (2002) Landscape structure, habitat fragmentation, and the ecology of insects, *Agricultural and Forest Entomology*, 4, 159-166.

Kullback S., Leibler R.A. (1951) On information and sufficiency, *The Annals of Mathematical Statistics*, 22, 79–86.

Luciano P. (1989) L'impiego delle trappole a feromone nella programmazione della lotta alla *Lymantria dispar* L., Proceedings of *Avversità del bosco e delle specie arboree da legno*, 15-16 October 1987, Italy, pp. 345-357.

Moilanen A. (1999) Patch occupancy models of metapopulation dynamics: efficient parameter estimation using implicit statistical inference, *Ecology*, 80, 1031-1043.

Sharov A.A., Colbert J.J. (1996) A model for testing hypotheses of gypsy moth, *Lymantria dispar* L., population dynamics, *Ecological Modelling*, 84, 31-51.

Tscharntke T., Brandl R. (2004) Plant-insect interactions in fragmented landscapes, *Annual Review of Entomology*, 49, 405–430.

Zhou G., Liebhold A.M. (1995) Forecasting the spatial dynamics of gypsy moth outbreaks using cellular transition models, *Landscape Ecology*, 10, 177-189.

Geostatistical modelling of regional bird species richness : exploring environmental proxies for conservation purpose

Giovanni Bacaro, Alessandro Chiarucci
BIOCONNET, BIOdiversity and CONservation NETwork, Dipartimento di Scienze, Ambientali"G. Sarfatti", Università di Siena, Via P. A. Mattioli 4, 53100, Siena, Italy.

Elisa Santi
IRPI-CNR, Via Madonna Alta 126, 06128 Perugia, Italy

Duccio Rocchini
Department of Biodiversity and Molecular Ecology, GIS and Remote Sensing Unit, Fondazione Edmund Mach, Research and Innovation Centre, Via E. Mach 1, 38010 S. Michele all'Adige, TN, Italy

Francesco Pezzo, Luca Puglisi
Centro Ornitologico Toscano, C.P. 470, 57100 Livorno, Italy

Abstract: Identifying spatial patterns in species diversity represents an essential task to be accounted for when establishing conservation strategies or monitoring programs. Predicting patterns of species richness by a model-based approach has recently been recognised as a significant component of conservation planning. Here, a spatially-explicit data-set on birds presence and distribution across the whole Tuscany region was analysed using geostatistical models. Species richness was calculated within 1x1 km grid cells and 10 environmental predictors were included in the analysis. A statistical model integrating spatial components of variation with predictive ecological factors of bird species richness was developed and used to obtain predictive regional maps of bird diversity hotspots.
Keywords: Bird richness, Conservation, Distribution maps, Natura 2000 Network, Predictive model, Semivariance, Spatial autocorrelation, Tuscany.

1. Introduction

The identification of spatial patterns in species diversity represents an essential task for biodiversity conservation strategies or monitoring programs. Recently, species distribution modeling emerged as a new approach to generate species distribution maps, on the basis of the relationship between species presence (or abundance) records and environmental variables. Typically, modeling methods attempt to predict the probability of occurrence of species as a function of a set of environmental variables. In particular, geostatistical modeling techniques, which have been developed mainly in the field of geography, are designed to model spatially dependent observations (Goovaerts 1997), but in recent years, such methodologies have been applied even in the ecological literature (Bacaro an Ricotta 2007). Birds are among the best-studied organisms, especially in Europe. They are often considered as excellent indicators of environmental changes and as good ecological proxies to assess the biodiversity values of an area. In this work, a geostatistical modelling approach was applied on the data provided by the "Monitoring Program of Breeding Birds of Tuscany", one of the most extensive regional bird monitoring programs in Italy. The aim of this paper is *i*) to describe the

spatial patterns of bird species richness and *ii*) to identify those environmental factors underlying these patterns. This latter point represents an important task in the ecological context since the environmental proxies driving bird richness could be used to decide conservation strategies.

2. Materials and Methods

Bird data: The bird species occurrence data were obtained from the Monitoring Program of Breeding Birds of Tuscany carried out by the Centro Ornitologico Toscano (www.centrornitologicotoscano.org) and based on Point Counts method (Bibby *et al.* 2000). Points were distributed according to a two stages sampling design: in randomly selected 10*10 km UTM cells, a number of 12-15 point counts were selected according to a second random sampling procedure. The original data set of geo-referenced observations was assembled to produce a regional map of bird species richness for cells of 1*1 km. Such a grid covering the whole Tuscany region resulted in 22060 cells, 3584 of which enclosed data on bird occurrences.

Putative determinants of bird species richness: for each 1*1 km cell, three sets of predictor/explanatory variables were derived and grouped according to a similarity criterion. I) Geographical features (4 predictors): the coordinates for each grid cell (Latitude and Longitude), elevation and distance to the sea were included in this group. II) Landscape feature and complexity (4 variables): Data on land cover were derived from the third level of the CORINE Land Cover Map. For each grid cell, the number of patches and the area (mean and standard deviation) covered by each land cover class was calculated. Landscape shape complexity was calculated by using the AWMSI (Area Weighted Mean Shape Index). The third level data of the CORINE Land Cover were used for calculating the Shannon index. III) Primary Productivity (2 variables): NDVI (Normalized Difference Vegetation Index) and its standar deviation were used on to discriminate between the amount of biomass characterising different vegetation types.

Geostatistical modelling: a combined multi-predictor model was developed in this study, and it was further used in conjunction with geostatistical techniques to predict birds diversity in 1 x 1 km grid cells across the whole Tuscany region. Statistical modelling process was organised into the following three parts: 1) Data transformation (normalization); 2) Building the generalized linear spatial model: once the response variable (number of bird species) at each grid cell within the Tuscany region was denoted as:

$$\{x_i, y_i : i = 1, \ldots, n\} \qquad (1)$$

where x_i identifies the spatial location (in two-dimensional space - longitude and latitude expressed in kilometres) and y_i is the bird richness value associated with the location x_i, a geostatistical (isotropic) model can be defined as:

$$Y_i = S(x_i) + Z_i : i = 1, \ldots, n \qquad (2)$$

where

$$\{S(x) : x \in \mathbb{R}^2\} \qquad (3)$$

is a Gaussian process with a spatially varying mean $\mu(x)$ defined by a classical linear regression model. The described Gaussian process is also characterized by a variance given by:

$$\sigma^2 = Var\{S(x)\} \qquad (4)$$

and by a positive-defined correlation function:

$$\rho(u) = Corr\{S(x), S(x')\} \quad (5)$$

defining the way correlation function decays to zero for increasing distances occurring between observations at locations *x* and *x'*. Explanatory variables for modelling the large-scale variation in bird diversity were chosen via a model selection technique (AIC). Secondly, the residuals from the model were examined for spatial correlation and a suitable family of correlations was chosen. The estimates of the parameters in the trend surface (model spatial component) were updated using the quasi-Newton optimisation function (Byrd et al. 1995) followed by maximum likelihood estimation of the parameters of the covariance function using the residuals. 3) Universal kriging was used to predict expected bird richness (and its variation) in each 1x1 km grid cell across the whole Tuscany Region.

3. Results

The number of bird species per cell grid was normalized using a Box-Cox power of 0.184. Only 4 predictors were included in the predictive model (Table 1). The intercept of the estimated spatial varying mean resulted highly significant and was, consequently, included in the model.

Table 1: Description of explanatory variables (and their associated coefficients) included after stepwise selection in the spatial varying mean component (*** p<0.001).

Trend parameters (spatial varying mean)	**Estimated Value**
Intercept	3.066^{***}
NDVI St.Dev.	0.811^{***}
H' index	0.104^{***}
Mean elevation	-0.001^{***}
Distance sea	$>0.001^{***}$
Spatial Parameters	
Nugget (τ^2)	0.147
Partial sill (σ^2)	0.270
Range (φ)	0.054
Practical Range	0.162
Normalisation parameter (Box-Cox power)	
lambda (λ)	0.184
Covariance Function Parameters (Matèrn)	
Order (k)	0.5 (exponential model)

The modeled spatial parameters highlighted that autocorrelation in bird richness value existed and strongly influenced the number of observed species. In particular, the practical range was reached after 16 km, indicating the absence of further correlative structure in data after this threshold (see Figure 1). Relatively to the covariance function used to model the empirical variogram, the k=0.5 parameter was selected (corresponding to fit an exponential theoretical variogram with respect to the observed data). Predicted values were significantly related with observed bird richness (R^2 = 0.448, $p < 0.001$). For comparison, a simple multiple regression model without the inclusion of the spatial component in the analysis, showed a lower R^2 value (R^2=0.15, $p<0.001$). Predicted bird richness (and its associated variance) across all the Tuscan region is shown in Figure 2.

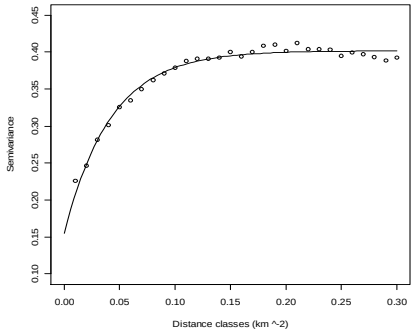


Figure 1: Plot of the empirical (circles) and fitted (solid line) semivariograms versus distance (km) obtained using the residuals after the spatial varying mean was subtracted by raw (normalized) data.

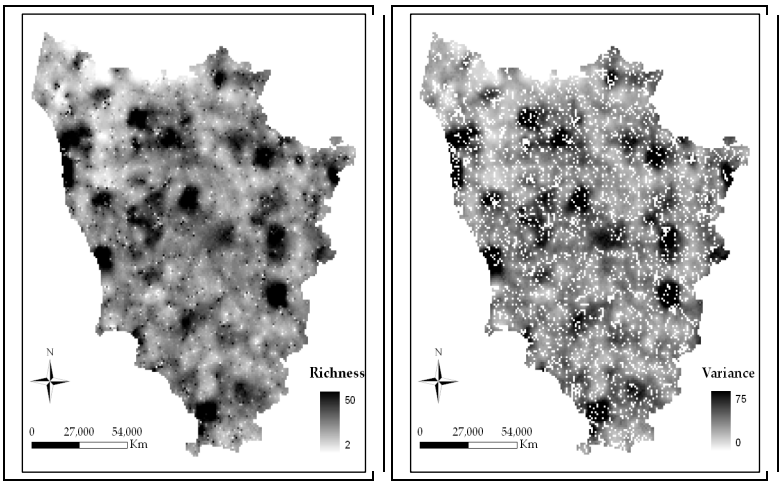


Figure 2: Regional pattern of bird species richness as expected under the described geostatistical model. a) Expected birds species richness and b) its expected variance.

4. Conclusions

By applying geostatistical models, a well-performing predictive model was obtained for the distribution of bird species richness in Tuscany by considering relatively few variables. Ancillary variables based on remotely sensed information (e.g., NDVI or Shannon *H'* derived from a classified image) can be used as powerful tools to model the spatial variation of bird species richness and locate biodiversity hotspots. Moreover, geostatistical models own the advantage to incorporate information of environmental co-variation and neighborhood effects, improving the quality of predictions.

References

Bacaro G, Ricotta C (2007) A spatially explicit measure of beta diversity. Community Ecol 8: 41-46.

Byrd RH, Lu P, Nocedal J, Zhu C (1995) A limited memory algorithm for bound constrained optimization.SIAM J. Scientific Computing16: 1190–1208.

Bibby CJ, Burgess ND, Hill DA, Mustoe SH (2000) Bird census techniques, 2nd edn. Academic Press, London.

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York.

Spatial Bayesian Modeling of Presence-only Data

Fabio Divino
S.T.A.T., University of Molise, fabio.divino@unimol.it

Natalia Golini, Giovanna Jona Lasinio
Department of Statistical Sciences, "Sapienza" University of Rome

Antti Penttinen
Department of Statistics, University of Jyväskylä

Abstract: When the only available information is the true presence of a species at few locations of a study area we refer to the data as *presence-only data*. Presence-only data problem can be seen as a missing data problem with asymmetric and partial information on a presence-absence process. This problem often characterizes ecological studies requiring the prediction of potential spatial extent of a species in suitable areas. Here we propose a Bayesian logistic spatial model adapted to presence-only data with environmental covariates available over the entire area. The spatial dependence among the observations is modelled indirectly as a latent Gaussian Markov field over the landscape, through a data augmentation MCMC algorithm we are able to estimate regression parameters jointly with the prevalence.

Keywords: Bayesian model, Data augmentation, MCMC, Presence-only data, Spatial distribution.

1 Introduction

In the environmental sciences, the evaluation of spatial distribution of species and its interaction with ecological variables is of primary interest *i.e.* to better plan and manage strategies in habitat conservation. When presence/absence information on a species is available in a given area together with environmental covariates, the logistic regression model represents the natural approach to estimate the prevalence of such species. Unfortunately, in many ecological studies, the collection of definitive absences can be expensive or difficult. In those cases the information available is not complete, we can observe only presences (Pierce and Boyce 2006) of the species at few locations jointly with the environmental covariates referred to the whole study area. In this work we propose a hierarchical Bayesian model to handle presence-only data, based on an adjusted logistic regression model (Ward et al. 2009). Following Divino et al. (2011) we introduce a random approximation of the correction factor

in the model that allows us to overcome the need to know *a priori* the prevalence of the species. We can estimate regression parameters jointly with prevalence through a data augmentation MCMC algorithm (Divino et al. 2011). We account for spatial variation adding a spatial random effect in the regression function.

2 Materials and Methods

With respect to a population $\mathcal{P}$ of spatially referenced sites i, let Y be a binary presence/absence process, X a set of covariates and $\mathcal{P}_p$ the subset of $\mathcal{P}$ where the species is present ($Y = 1$). When only presences are observed, samples (S_p) from the process Y can be drawn only from the population $\mathcal{P}_p$ and the usual case-control approach in logistic regression cannot be adopted as absences ($Y = 0$) are not directly observed. Lancaster and Imbens (1996) and Ward et al. (2009) proposed to overcome this problem by considering a completed sample composed by S_p and a second sample S_u, independent of S_p, ideally taken from the whole population $\mathcal{P}$. In this way the complete data sample S is composed by n_p presences (observed in S_p) and n_u unobserved values (S_u). Let Z be a stratum variable such that $Z_i = 0$ if $i \in S_u$ and $Z_i = 1$ if $i \in S_p$. Notice that $Z_i = 1$ implies $Y_i = 1$ while $Z_i = 0$ implies that Y_i can assume value in $\{0, 1\}$. Hence we can identify the following quantities: $(Z = 0, Y = 0)$ n_{0u} is the unknown number of absences in the subsample S_u, $(Z = 0, Y = 1)$ n_{1u} is the unknown number of presences in the subsample S_u, $(Z = 1, Y = 1)$ n_{1p} is the number of observed presences in the subsample S_p, n_0 is the unknown total number of absences in S, n_1 is the unknown total number of presences in S and $n = n_1 + n_0$ is the complete sample size. All the unknowns are random quantities induced by a censoring effect acting on the complete sample S. In particular we can write n_{1u} as $\tilde{n}_{1u} = \sum_{i \in S_u} Y_i$, where the $\sim$ represents the random nature of the quantity. Now let $\pi = P(Y = 1)$ be the prevalence of the species in the area, under the assumption that S_u is a random sample from the population $\mathcal{P}$ we have that $E[\tilde{n}_{1u}] = \pi n_u$. If we assume that the covariates X, concerning the environmental information on the process Y, are available for all sites in the population, we can use the approach introduced by Ward et al. (2009) and developed in a Bayesian framework by Divino et al. (2011). For a generic site in the sample with covariates x, starting from the usual case-control logistic model the conditional probability that a species of interest is present is given by

$$P(Y = 1|s = 1, \eta; x) = \frac{\exp\{\eta(x) + \log(\frac{\gamma_1}{\gamma_0})\}}{1 + \exp\{\eta(x) + \log(\frac{\gamma_1}{\gamma_0})\}} \tag{1}$$

where $s = 1$ denotes that the site is included in S, $\eta(x)$ is the regression function, $\gamma_0 = P(s = 1|Y = 0)$ and $\gamma_1 = P(s = 1|Y = 1)$ are the unknown probabilities of sampling from the absences and from the presences respectively. The ratio $\frac{\gamma_1}{\gamma_0}$ adjusts the logistic model under the case-control design. Following Ward et al. (2009), we can manage the presence-only data problem by considering the joint probability

distribution of Y and Z and write the full likelihood model (see Ward et al. 2009 for details). We can also consider the observed likelihood, built only with respect to the stratum variable Z that results in an average over the process Y. In both likelihood models, the unknown ratio $\frac{\gamma_1}{\gamma_0}$ can be approximated as follow:

$$\frac{\tilde{\gamma_1}}{\tilde{\gamma_0}} \approx \frac{\tilde{n}_{1u} + n_p}{\tilde{n}_{1u}} \tag{2}$$

the above expression can be handled by a data augmentation step in the estimation procedure. The regression function adopted in this work is linear with a spatially structured random effect u accounting for latent factors introducing geographical dependence into species distribution. We can now write the hierarchical Bayesian model. Let δ be the vector of hyperparameters with hyperprior $p(\delta)$. Conditioned on δ, the regression parameters, β, are Gaussian random variables and the random effect u is a Gaussian Markov random field. Given β, u and the covariate x, the process Y is set of Bernoulli random variable with probability of occurence $\pi_s(x) = P(Y = 1|s = 1, \eta; x)$. At the lowest level of the hierarchy, the conditional distribution of Z given Y can be easily derived from the above described relations between the two processes. Then, the hierarchical Bayesian model is given by: (i) $\delta \sim p(\delta)$; (ii) $\beta|\delta \sim MN(\delta)$ and $u|\delta \sim GMRF(\delta)$; (iii) $Y_i|s_i = 1, \beta, u_i, x_i \sim Be[\pi_s(x_i)]$; (iv) $Z_i|Y_i, s_i \sim P(Z_i|Y_i, s_i = 1)$. Notice that the spatial structure of the random effect u is given by the geographical neighborhood system among all sites in the population $\mathcal{P}$. In the following scheme we describe the MCMC algorithm implementing the estimation of our model:

Step 0: initialize δ, β, u and Y over $\mathcal{P}$;
Step 1: set $n_{1u} = \sum_{i \in S_u} Y_i$;
Step 2: sample $\delta \sim P(\delta|Y, Z, \beta, u)$;
Step 3: sample $\beta \sim P(\beta|Y, Z, \delta)$;
Step 4: sample $u \sim P(u|Y, Z, \delta)$ over $\mathcal{P}$;
Step 5: sample $Y_i \sim P(Y_i|Z, \beta, u_i, x_i)$ over $\mathcal{P}$.

Remark that we need to perform data augmentation (Step 4 and Step 5) over the entire population $\mathcal{P}$ for both u and Y processes in order to consider the spatial structure of the sites enclosed in both samples S_u and S_p. The only requirement to perform the augmentation is that the covariates X are available for all sites in $\mathcal{P}$. A nice feature of this estimation procedure is that we can easily obtain the prevalence estimate $\hat{\pi}_u = \frac{\bar{n}_{1u}}{n_u}$, where $\bar{n}_{1u}$ is the MCMC average of samples drawn in Step 1.

3 Results

In this section we report preliminary results from a small simulation study aiming at investigating the behaviour of our proposal in a very simple situation. We generate a population of 100 observations on a regular 10×10 lattice from the above described model. In this example Y is obtained from the logistic model $\eta(\boldsymbol{X}) = \beta x + u$, where

$\beta = -2$, the covariate X is generated from a mixture distribution with two Gaussian components with standard deviation $\sigma_1 = \sigma_2 = 0.5$ and mean $\mu_1 = -2$ and $\mu_2 = 2$, u is a zero mean intrinsic first order Gaussian Markov random field with precision $k = 1.5$ and prevalence $\pi = 0.1$. From this population we obtain 100 samples by randomly thinning 30% of the available presences. We compare the performance of our model (M1), with unknown prevalence, with the same model but with known prevalence in the logistic correction (M2) and with the non spatial model prosed in Divino et al. (2011) (M3). The three models are fitted with the same prior settings: $\beta \sim N(0, 100)$ and k fixed (for M1 and M2). We run 20000 iterations of the MCMC procedure with a burn-in of 10000. To evaluate models performances we compute 95% credibility intervals (CI) for β in each simulation using the 10000 samples from the posterior distribution, the same intervals for the prevalence are computed from the 100 simulations and the misclassification error is computed for each model by setting to 1 grid cells with occurrence probability larger then 0.5 and compare results with the "true" population. Results are as expected: the "best" model in terms of point estimates accuracy is M2 with smaller CI for $\hat{\beta}$ and $\hat{\pi}$, followed by M1; all models have a tendency to overfit with empirical coverage around 99%. In terms of predictive capacity the average misclassification error is around 3% for all models, as expected M1 and M2 better perform as far as the localization of presences is concerned.

4 Concluding remarks

The above preliminary results are encouraging, especially in terms of predictive capacity of the proposed model. Several issues will be object of further work, such as identifiability problems related to a not zero intercept. Extensive simulation studies will be carried on too.

References

Divino F., Jona Lasinio G., Golini N., Pettinen A. (2011) Data Augmentation Approach in Bayesian Modelling of Presence-only Data, *Procedia Environmental Sciences* to appear.

Lancaster T., Imbens G. (1996). Case-control studies with contaminated controls. *Journal of Econometrics* 71,145-160.

Pearce J.L, Boyce, M.S. (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology* 43: 405-412

Ward G, Hastie T, Barry S, Elith J, Leathwick A. (2009) Presence-only data and the EM algorithm. *Biometrics*; 65: 554-563.

The deep-water rose shrimp in the Ionian Sea: a spatio-temporal analysis of zero-inflated abundance data

D'Onghia G., Maiorano P., Carlucci R., Tursi A.,
Dipartimento di Biologia, Università degli Studi di Bari "Aldo Moro"
g.donghia@biologia.uniba.it

Pollice A., Ribecco N., Calculli C., Arcuti S.,
Dipartimento di Scienze Statistiche "Carlo Cecchi",
Università degli Studi di Bari "Aldo Moro"

Abstract: In the ecological field, the sampling of abundance data is often characterized by the zero inflation of population distributions. Constrained zero-inflated GAM's (COZIGAM) are obtained assuming that the probability of non-zero inflation and the mean non-zero-inflated population abundance are linearly related. Models of this class have been applied to a spatio-temporal case study concerning the deep-water rose shrimp, *Parapenaeus longirostris* (Lucas, 1846). Abundance data were collected during 16 experimental trawl surveys conducted from 1995 to 2010 in the Ionian Sea. The sampling design adopted was random-stratified by depth, with proportional allocation of hauls to the area of each depth range and geographical sector. Density index (N/km^2) and length (mm) were considered for each haul identified by time, depth, geographic coordinates and geographical sector.

Keywords: Zero-inflated data, COZIGAM, GAM, density, size, *Parapenaeus longirostris.*

1. Introduction

In the ecological field, the sampling of abundance data is often characterized by the zero inflation of population distributions. Many Mediterranean species show such a distribution due to their adaptation to the variable environmental conditions. One of these is the deep-water rose shrimp, *Parapenaeus longirostris* (Lucas, 1846), widespread throughout the whole Mediterranean Sea at depths between 20 and 700 m. The Ionian Sea is a basin where this shrimp represents the bulk of the catch due to the trawl fishing carried out on between shelf edge and upper slope. Aspects of the distribution and population biology of this shrimp are reported in D'Onghia et al. (1998) and Abelló et al. (2002). Its spatio-temporal distribution in the Ionian Sea for the period 1995-2010 has been studied and the relevant results have been reported in this paper.

	Density		***Length***	
parameters	*estimate*	*p-value*	*estimate*	*p-value*
intercept			14.248	<0.000
depth$_f$(0,200]	5.240	<0.000		
depth$_f$(200,500]	5.348	<0.000		
depth$_f$(>500)	3.5254	<0.000		
depth			0.029	<0.000
alpha	- 1.030	<0.000		
delta1	1.004	<0.000		
delta2	0.666	<0.000		
smooth terms	*df*	*p-value*	*df*	*p-value*
s(lon, lat)	28.855	<0.000	18.111	0.005
s(year)	8.507	<0.000	7.091	<0.000

Table 1: COZIGAM's estimates for the density index, GAM's estimates for the length.

2. Materials and Methods

Abundance data were collected during 16 experimental trawl surveys conducted from 1995 to 2010 in the Ionian Sea as part of the international MEDITS project funded by EC (Bertrand et al., 2000). The samples analyzed come from a total of 1052 hauls carried out during day-light hours between 10 and 800 m in the spring season (May-June). The sampling design adopted was random-stratified by depth, with proportional allocation of hauls to the area of each depth range and geographical sector. Density index (N/km^2) and carapace length (mm) were considered for each haul identified by time, depth, geographic coordinates and geographical sector.
A general approach to zero-inflated data modeling consists in assuming the response distribution as a probabilistic mixture of a zero and a non-zero generating process. Zero-inflated general linear models (ZIGLM) can be readily extended to include smooth effects of covariates giving rise to ZIGAMs. A constrained zero-inflated GAM (COZIGAM) is obtained assuming that the probability of non-zero inflation and the mean non-zero-inflated population abundance are linearly related. In this paper an analysis of the density index based on COZIGAM's is proposed. As *P. longirostris* carapace length is not affected by zero-inflation, given that no measurements are available when the density index is null, this variable is analyzed in the GAM's framework. The R libraries `COZIGAM` (Liu and Chan, 2010) and `mgcv` (Wood, 2006) were used for the data analysis.

3. Results

Preliminary exploratory data analysis (not reported) showed a discontinuous higher presence of zeroes and small density values at lower (shallower than 200 m) and higher (deeper than 500 m) depths. This lead to considering the factorization of the depth variable accordingly in the model for the density index. We propose the following specification for the mean of the log-Gaussian non-zero generating process:

$$\mu = s(\text{lon,lat}) + s(\text{year}) + \text{depth}_f$$

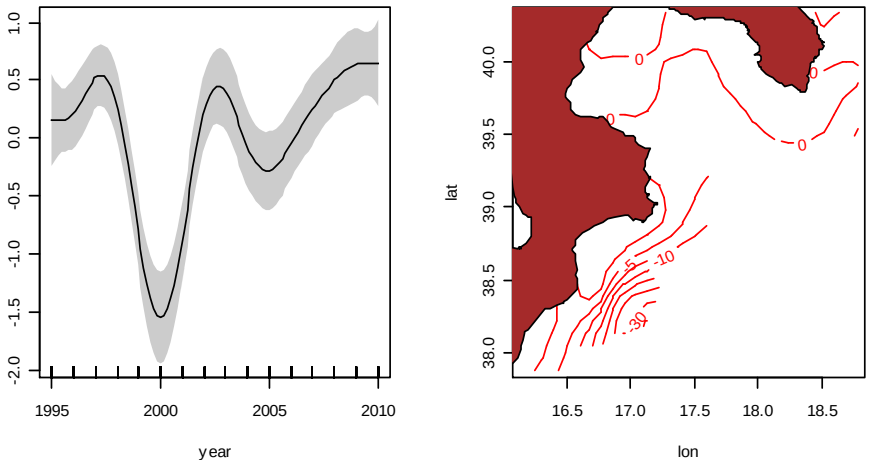


Figure 1: COZIGAM estimated effects of space and time for the *P. longirostris* density.

and assume that the smooth spatial effect and the temporal one have a different importance on the non-zero inflation probability, according to the following proportionality constraint:

$$logit(p) = \alpha + \delta_1 s(\text{lon,lat}) + \delta_2 s(\text{year}) + \text{depth}_f$$

In Tab. 1 we report the estimates of the COZIGAM model effects for the density index. The estimated effects of the three depth levels agree with the observed data. The estimates of δ_1 and δ_2 have significantly positive values, showing that the zero inflation probability decreases with the density value.

In Fig. 1, Left we report the estimate of the smooth temporal effect showing a severe drop of the density index in 2000. This was also expected according to the results of the preliminary exploratory data analysis which described a decreasing in the density index in the 1999-2001 years. The map of the spatial effect (Fig. 1, Right) reveals a wide distribution of the species along the Ionian arc with three main areas with a greater density.

A Gaussian GAM for the *Parapenaeus* carapace length is specified as follows:

$$\mu = intercept + s(lon,lat) + s(year) + depth$$

In this case the exploratory data analysis shows a continuous linear relation with the depth, leading to consider the unfactorized variable within a linear term; carapace lengths increase with deeper sea beds (Tab. 1). The smooth estimated temporal effect (Fig. 2, Left) has an opposite behavior with respect to the density index, showing a peak in 2000. Also a second peak in 2005 and a drop in 2008 are noticeable. The former could be due to a lower density of greater specimens, the latter could be in relation to an increase of juveniles in the sampled population. The map of the spatial effect shows greater sizes in the Gallipoli (Apulia) and Roccella Ionica (southern Calabria) fishery districts (Fig. 2, Right).

4. Concluding remarks

These results confirm the knowledge on density and size distribution of *P. longirostris* in the Ionian Sea (D'Onghia et al., 1998; Abello et al. 2002) revealing geographic and temporal effect both on density and size. The increasing density together with the

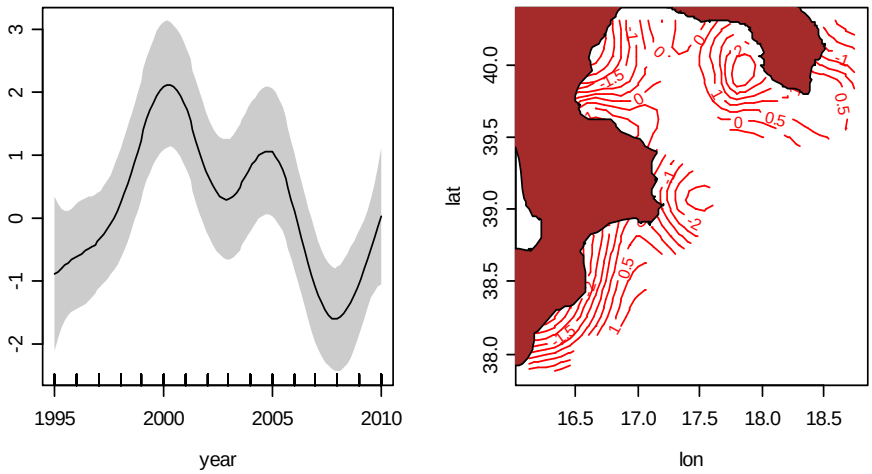


Figure 2: GAM estimated effects of space and time for *P. longirostris* carapace length.

decreasing size observed in some years could be related to the increase in the recruitment detected for the deep-water rose shrimp. This will require further investigation in order to identify the environmental variables affecting the changes observed in the species distribution.

References

Abelló P., Abella A., Adamidou A., Jukic-Peladic S., Maiorano P., Spedicato M. T. (2002) Geographical patterns in abundance and population structure of Nephrops norvegicus and Parapenaeus longirostris (Crustacea: Decapoda) along the European Mediterranean coasts, *Scientia Marina*, 66 (Suppl. 2), 125-141.

Bertrand J.A., Gil de Sola L., Papaconstantinou C., Relini G., Souplet A. (2000) An international bottom trawl survey in the Mediterranean: the MEDITS programme, *IFREMER Actes de Colloques*, 26, 76-93.

D'Onghia G., Matarrese A., Maiorano P., Perri F. (1998) Valutazione di Parapenaeus longirostris (Lucas, 1846) (Crustacea, Decapoda) nel Mar Ionio. *Biologia Marina Mediterranea*, 5 (2), 273-283.

Liu H., Chan K. S. (2010) Robust Introducing COZIGAM: An R Package for Unconstrained and Constrained Zero-Inflated Generalized Additive Model Analysis, *Journal Statistical Software*, 35(11), 1-26.

Liu H., Ciannelli L., Decker M. B., Ladd C., Chan K. S.. (2010) Nonparametric Threshold Model of Zero-Inflated Spatio-Temporal Data with Application to Shifts in Jellyfish Distribution, *Journal of Agricultural, Biological, and Environmental Statistics*.

Wood S. N. (2006). *Generalized Additive Models: An Introduction with R*, Chapman & Hall/CRC.

A few links between the notion of Entropy and Extreme Value Theory in the context of analyzing climate extremes
[1]

Philippe Naveau
Laboratoire des Sciences du Climat et l'Environnement (LSCE) CNRS, France,
naveau@lsce.ipsl.fr

Théo Rietsch
LSCE and Université de Strasbourg

Armelle Guillou
Université de Strasbourg

James Merleau
IREQ Canada

Abstract:
Climate change could have an important impact on the distribution of future extreme events. To assess such changes, it is essential to develop statistical tools based on Extreme Value Theory. In this talk, we make and study some connections between the notion of entropy (divergence) and Extreme Value Theory. We apply these links to detect changes in extremes.

Keywords: Extreme Value Theory, entropy, climate

Materials and Methods

The Kullback-Leibler information (Kullback, 1968) is defined as

$$I(f;g) = E_f \left\{ \log \left(\frac{f(\mathbf{Z})}{g(\mathbf{Z})} \right) \right\}, \tag{1}$$

and it measures the entropy distance (Robert, 2001) between the probability densities f and g for a random variable $\mathbf{Z}$. Kullback (1968) also refers to this quantity as the directed divergence to distinguish it from the divergence given by

$$J(f;g) = I(f;g) + I(g;f), \tag{2}$$

[1]Part of this work has been supported by the EU-FP7 ACQWA Project (www.acqwa.ch) under Contract Nr 212250, by the PEPER-GIS project, by the ANR-MOPERA project, by the ANR-McSim project and by the MIRACCLE-GICC project.

which is a symmetrical measure relative to f and g. This notion has been exhaustively used and studied in many research fields. Here we explore this concept within the framework of climatology and Extreme Value Theory (EVT). While the divergence (2) is expressed in function of densities, it is more convenient to work the tail distribution when analyzing large excesses. In this talk, we propose an approximation to bypass the need of computing densities. This allows us to derive and study new estimators of the entropy. We apply this approach to the important problem of detecting changes in our warming climate.

References

[Kullback, 1968] Kullbcak, S. (1968). *Information Theory and Statistics*, 2nd ed. New York: Dover.

[Resnick, 2007] Resnick, S. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Operations Research and Financial Engineering, Springer, New York.

[Robert, 2001] Robert, C. P. (2001). *The Bayesian Choice*. New York: Springer-Verlag.

Geoadditive modeling for extreme rainfall data

Chiara Bocci, Alessandra Petrucci
Department of Statistics "Giuseppe Parenti", University of Firenze,
bocci@ds.unifi.it, alessandra.petrucci@unifi.it

Enrica Caporali
Department of Civil and Environmental Engineering, University of Firenze,
enrica.caporali@unifi.it

Abstract: Extreme value models and techniques are widely applied in environmental studies to define protection systems against the effects of extreme levels of environmental processes. Regarding the matter related to the climate change science, a certain importance is cover by the implication of changes in the hydrological cycle. Among all hydrologic processes, rainfall is a very important variable as it is a fundamental component of flood risk mitigation and drought assessment, as well as water resources availability and management. We implement a geoadditive mixed model for extremes with a temporal random effect assuming that the observations follow generalized extreme value distribution with spatially dependent location. The analyzed territory is the catchment area of Arno River in Tuscany in Central Italy.

Keywords: GEV distribution, geoadditive mixed model, hydrologic processes

1 Introduction

Environmental extreme events such as floods, earthquakes, hurricanes, may have a massive impact on everyday life for the consequences and damage that they cause. For this reason there is considerable attention in studying, understanding and predicting the nature of such phenomena and the problems caused by them, not least because of the possible link between extreme climate events and climate change. A number of theoretical modeling and empirical analyses have also suggested that notable changes in the frequency and intensity of extreme events, including intense rainfall and floods, may occur even when there are only small changes in climate (Katz and Brown, 1992).

In this framework, in the past two decades there has been an increasing interest for statistical methods that model rare events (Coles, 2001). The Generalized Extreme Value distribution (GEV) is widely adopted model for extreme events in the univariate context. For modeling extremes of non-stationary sequences it is commonplace to use the GEV as a basic model, and to handle the issue of non-stationarity by regression modeling of the GEV parameters.

Here we implement a geoadditive mixed model for extremes with a temporal random effect. We assume that the observations follow a generalized extreme value distribution whose locations are spatially dependent where the dependence is captured using the geoadditive model. The analyzed territory is the catchment area of Arno River in Tuscany in Central Italy.

2 Materials and Methods

The investigation is developed on the catchment area of Arno River almost entirely situated within Tuscany, Central Italy. The time series of annual maxima of daily rainfall recorded in 415 rain gauges are analyzed. In order to have enough rain gauges observations to estimate both the spatial component and the year specific effect, we reduce the time series length to the post Second World War period and we consider only stations with at least 30 hydrologic years of data, even not consecutive. The final dataset is composed by the data recorded from 1951 to 2000 at 118 rain gauges for a total of 4903 observations.

Recently to handle the issue of non-stationarity of the GEV parameters, Padoan and Wand (2008) discuss how generalized additive models (GAM) with penalized splines can be carried out in a mixed model framework for the GEV family.

Geoadditive models, introduced by Kammand and Wand (2003), are a particular specification of GAM that models the spatial distribution of y with a bivariate penalized spline on the spatial coordinates. Suppose to observe n sample maxima y_{ij} at spatial location $\mathbf{s}_{ij}$, $\mathbf{s} \in \mathbb{R}^2$, $j = 1, \ldots, p$ and at time $i = 1, \ldots, t$. In order to model both the spatial and the temporal influence on the annual rainfall maxima, we consider a geoadditive mixed model for extremes with a temporal random effect:

$$\begin{cases} y_{ij}|s_{ij} \sim \text{GEV}(\mu\left(s_{ij}\right), \psi, \xi) \\ \mu\left(\mathbf{s}_{ij}\right) = \beta_0 + \mathbf{s}_{ij}^T \boldsymbol{\beta}_s + \sum_{k=1}^{K} u_k b_{tps}(\mathbf{s}_{ij}, \boldsymbol{\kappa}_k) + \gamma_i, \end{cases} \tag{1}$$

where μ, ψ and ξ are respectively location, scale and shape parameters of the GEV distribution, b_{tps} are the low-rank thin plate spline basis functions with K knots and γ_i is the time specific random effect. The model (1) can be written as a mixed model

$$\mathbf{y}|\left(\mathbf{u}, \boldsymbol{\gamma}\right) \sim \text{GEV}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{D}\boldsymbol{\gamma}, \psi, \xi). \tag{2}$$

with

$$\text{E}\begin{bmatrix}\mathbf{u} \\ \boldsymbol{\gamma}\end{bmatrix} = \begin{bmatrix}0 \\ 0\end{bmatrix}, \qquad \text{Cov}\begin{bmatrix}\mathbf{u} \\ \boldsymbol{\gamma}\end{bmatrix} = \begin{bmatrix}\sigma_u^2 \mathbf{I}_K & 0 \\ 0 & \sigma_\gamma^2 \mathbf{I}_t\end{bmatrix}.$$

where

$$\boldsymbol{\beta} = \left[\beta_0, \boldsymbol{\beta}_s^T\right] \qquad \mathbf{u} = [u_1, ..., u_K] \qquad \boldsymbol{\gamma} = [\gamma_1, ..., \gamma_t]$$
$$\mathbf{X} = \left[1, \mathbf{s}_{ij}^T\right]_{1 \leq ij \leq n} \qquad \mathbf{D} = [d_{ij}]_{1 \leq ij \leq n}$$

with d_{ij} an indicator taking value 1 if we observe a rainfall maxima at rain gauge j in year i and 0 otherwise, and $\mathbf{Z}$ the matrix containing the spline basis functions, that is

$$\mathbf{Z} = [b_{tps}(\mathbf{s}_{ij}, \boldsymbol{\kappa}_k)]_{1\leq ij\leq n, 1\leq k\leq K} = [C(\mathbf{s}_{ij} - \boldsymbol{\kappa}_k)]_{1\leq ij\leq n, 1\leq k\leq K} \cdot [C(\boldsymbol{\kappa}_h - \boldsymbol{\kappa}_k)]^{-1/2}_{1\leq h,k\leq K},$$

where $C(\mathbf{v}) = \|\mathbf{v}\|^2 \log \|\mathbf{v}\|$ and $\boldsymbol{\kappa}_1, ..., \boldsymbol{\kappa}_K$ are the spline knots locations.

3 Results

The geoadditive mixed model for extremes (2) can be naturally formulated as a hierarchical Bayesian model and estimated under the Bayesian paradigm. Following the specifications of Padoan (2008), our complete hierarchical Bayesian formulation is

$$\begin{aligned}
&\text{1st level} \quad y_i | (\mathbf{u}, \boldsymbol{\gamma}) \overset{\text{ind}}{\sim} \text{GEV}([\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{D}\boldsymbol{\gamma}]_i, \psi, \xi) \\
&\text{2st level} \quad \mathbf{u}|\sigma_u^2 \sim N(0, \sigma_\gamma^2 \mathbf{I}_K) \qquad \boldsymbol{\gamma}|\sigma_\gamma^2 \sim N(0, \sigma_\gamma^2 \mathbf{I}_t) \qquad \boldsymbol{\beta} \sim N(0, 10^4 \mathbf{I}) \\
&\qquad\qquad \xi \sim \text{Unif}(-5, 5) \qquad \psi \sim \text{InvGamma}(10^{-4}, 10^{-4}) \\
&\text{3st level} \quad \sigma_u^2 \sim \text{InvGamma}(10^{-4}, 10^{-4}) \qquad \sigma_\gamma^2 \sim \text{InvGamma}(10^{-4}, 10^{-4}).
\end{aligned}$$

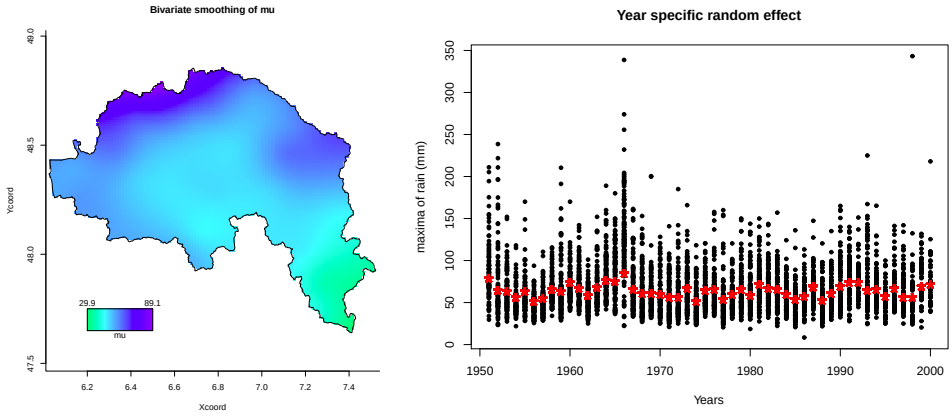
where the parameters setting of the priors distributions for ξ, ψ, $\boldsymbol{\beta}$, σ_u^2, σ_γ^2, corresponds to non-informative priors.

Given the complexity of the proposed hierarchical models, we employ `OpenBUGS` Bayesian MCMC inference package to do the model fitting. We access `OpenBUGS` using the package `BRugs` in the `R` computing environment. We implement the MCMC analysis with a burn-in period of 40000 iterations and then we retain 10000 iterations, that are thinned by a factor of 5, resulting in a sample of size 2000 collected for inference. Finally, the last setting concern the thin plate spline knots that are selected setting $K = 30$ and using the *clara* space filling algorithm of Kaufman and Rousseeuw (1990), available in the `R` package `cluster`.

The resulting spatial smoothing component and time specific component of $\mu(s_{ij})$ are presented in Figures 1(a) and 1(b). Observing the map, it is evident the presence of a spatial trend in the rainfall extreme dynamic, even after controlling for the year effect. The spline seems to capture well the spatial dependence as it produce the same same patter of the Average Total Annual Precipitation. The time influence is pointed out by the estimated year specific random effects, that present a strong variability through years.

4 Conclusions

We have implemented a geoadditive modeling approach for explaining a collection of spatially referenced time series of extreme values. We assume that the obser-



(a) (b)

Figure 1: Estimated spatial component (a) and year specific random effects (b) of $\mu(s_{ij})$. Black dots indicate the observed values.

vations follow generalized extreme value distributions whose locations are spatially dependent.

The results show that this model allows us to capture both the spatial and the temporal dynamics of the rainfall extreme dynamic.

Under this approach we expect to reach a better understand of the occurrence of extreme events which are of practical interest in climate change studies particularly when related to intense rainfalls and floods, and hydraulic risk management.

References

Coles, S.G. (2001) *An Introduction to Statistical Modeling of Extreme Values*, Springer, London.

Kammann, E.E., Wand, M.P. (2003) Geoadditive models, *Applied Statistics*, 52, 1-18.

Katz R.W., Brown B.G. (1992) Extreme events in a changing climate: variability is more important than averages, *Climate Change*, 21, 289-302.

Kaufman, L., Rousseeuw, P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.

Padoan, S.A. (2008) *Computational methods for complex problems in extreme value theory*, Ph.D. thesis, Ph.D. in Statistical Science, Department of Statistical Science, University of Padova.

Padoan, S.A., Wand, M.P. (2008) Mixed model-based additive models for sample extremes, *Statistics and Probability Letters*, 78, 2850-2858.

Spatio-temporal rainfall trends in southwest Western Australia

Ken Liang
University College London, ken@stats.ucl.ac.uk

Richard Chandler
University College London

Giampiero Marra
University College London

Abstract: Over the past several decades, there have been significant reductions in rainfall across southwest Western Australia. In the present work, the spatial and temporal structure of these reductions are investigated using generalized additive models. This involves smoothing over both space and time, to allow spatio-temporal interactions, as well as allowing for spatial correlation to ensure that standard errors are constructed appropriately for inference. The proposed method is computationally convenient as models are fitted as though different spatial locations are independent, and inference is subsequently adjusted for inter-site dependence. The results quantify precisely the spatially-varying nature of the decreasing rainfall trends.

Keywords: Spatio-temporal modelling, generalized additive models, tensor product smooth.

1 Introduction

Several decades of below average rainfall combined with a noticeable shift toward drier winter conditions, has focused attention on water resource availability and agricultural management in southwest Western Australia (SWWA) (Bates et al., 2008). The aim of this analysis is to characterize spatio-temporal trends in rainfall intensity and occurrence across SWWA. This is achieved by fitting generalized additive models (GAM) to data from selected locations in the study area. A key issue here is to take due account of potential spatial and temporal correlations in the data. Our approach to this is to treat the data as independent during fitting and subsequently to adjust standard errors for the dependence. This provides a computationally convenient means of addressing the problem.

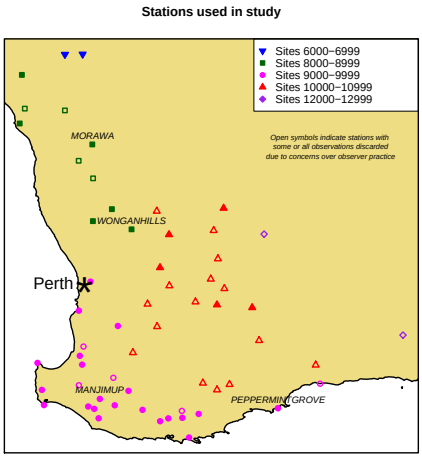


Figure 1: Map of SWWA indicating the stations used in the analysis

2 Materials and Methods

Daily rainfall readings in millimetres, from 60 weather stations (see Figure 1) for the period 1940-2010 have been used, although not all sites were operational throughout this period and some had missing observations. We consider the data for the winter months only, from May to July, which accounts for most of the region's annual rainfall. The daily rainfall data is aggregated in two ways, the proportion of wet days and total rainfall on wet days, for each site and year.

Prior to fitting the model, the daily rainfall data were subject to data quality checks. Issues such as inconsistencies in the data related to observer practice across the different sites, differences in the resolution of the observation recordings, and thresholding the data prior to analysis to ensure consistency have all been considered. These problems are typical in the rainfall modelling literature, the interested reader is referred to the results discussed in Yang et al. (2006) and Chandler et al. (2011). We shall not go into these details here.

Rainfall, particularly at the daily time scale, typically displays some form of temporal dependence, however at annual timescales they are relatively independent. Here, the winter rainfall is aggregated at an annual level, it seems reasonable to proceed in the first instance as though observations are independent between years. Since interest lies in characterizing temporal trends which may have a complex structure and may be spatially-varying, we adopt a nonparametric approach and represent the spatio-temporal trend surface as a smooth three-dimensional function of space and time:

$$E[y_{it}] = f(longitude_i, latitude_i, time_t),$$

where y_{it} is the aggregated annual winter rainfall at location $i = 1, \ldots, 60$ and year $t = 1, \ldots, 71$ and is assumed to be normally distributed. To account for different number of observations per year at each station, when fitting our model each observation is weighted by the number of contributing daily values. Our method uses the spline framework for nonparametric function estimation. To model smooths of several variables, when the variables are on different scales (the units of time (years) and space (km) are different), tensor product scale invariant smooths are required. Separate smoothing penalties are calculated for the three covariates so that the degree of smoothness is not necessarily the same for each covariate. All statistical analyses were done using the mgcv package (Wood, 2006) in the R software (R Development Core Team, 2010).

A key issue with this study is the spatially correlated nature of the data. Supposing that some assumption of spatial stationarity holds, the residuals from the fitted model were used to estimate the spatial correlation parameters obtained from the chosen variogram or correlation model, for the construction of the spatial correlation matrix. Specification of a covariance structure based on the spatial correlation of residuals ensures that the results are adjusted for spatial correlation. Because the focus of our approach is on estimating the mean function, not the correlation function, a very precise estimate of the latter is not required, and simple variogram models like the exponential will often suffice. The fitted function can be written as a linear smoother, $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ where $\mathbf{S}$ is the smoothing matrix. For spatially correlated observations, the true variance matrix is not diagonal. The model based variance matrix, $V(y)$, is replaced by the robust variance matrix, $\mathrm{Var}(y) = \mathbf{A^{1/2}RA^{1/2}}$ where $\mathbf{A}$ is a diagonal matrix, with the variance function $V(\mu)$, along diagonal elements and $\mathbf{R}$ is the spatial correlation matrix. Once the variance-covariance matrix is calculated, standard errors are then constructed in the usual way, as the square-roots of the elements on the main diagonal.

3 Results

After addressing spatial correlations, new 95% uncertainty bands were obtained which are now slightly wider than the unadjusted ones (see Figure 2). Interestingly, the width of the bands for the selected four sites differ significantly. This could be due the sparseness of data points at particular locations within SWWA which makes it difficult to characterize precipitation trends reliably. These declines were most pronounced in north and east of the study region, less so along the south coast.

4 Concluding remarks

To sum up, the used GAM framework enables us to appropriately incorporate all relevant covariates of space and time. In addition, accounting for the spatial dependence structure by assuming no correlation when fitting models and then adjusting

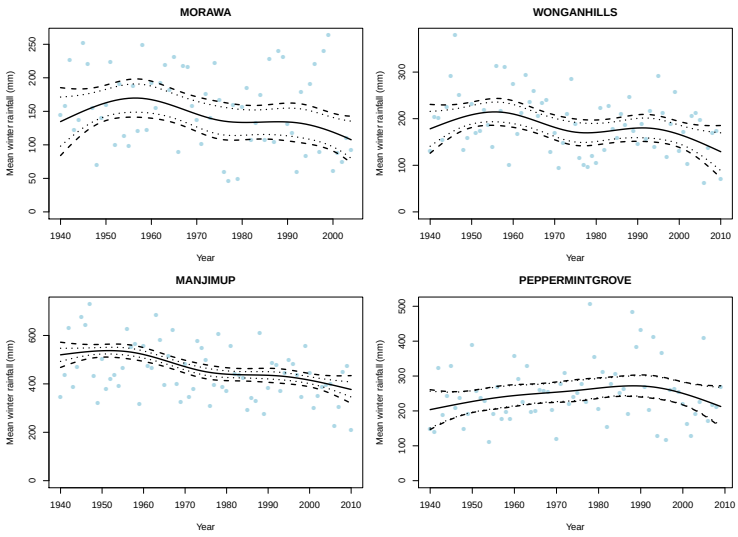


Figure 2: Fitted smooth curves (solid line) with unadjusted (dotted line) and adjusted (dashed line) 95% uncertainty bands for four selected sites in the SWWA region

the standard errors of estimates enable valid inferences that is robust. The results quantify precisely the spatially-varying nature of the decreasing rainfall trends.

References

Bates B., Hope P., Ryan B., Smith I. & Charles S. (2008) Key findings from the Indian Ocean Climate Initiative and their impact on policy development in Australia, *Climatic Change*, 89, 339-354.

Chandler R. E., Bates B. C. & Charles S. P. (2011) Rainfall trends in southwest Western Australia, in: *Statistical Methods for Trend Detection in the Environmental Sciences*, In Chandler R. E. & E. M. Scott (Eds.), Wiley, Chichester, 307-332.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing, ISBN 3-900051-07-0.

Wood, S. (2006) *Generalized Additive Models. An Introduction With R*, Boca Raton: Chapman & Hall/CRC.

Yang C., Chandler R. E., Isham V. S. & Wheater H. S. (2006) Quality control for daily observational rainfall series in the UK, *Water and Environment Journal*, 20(3), 185-193.

Stochastic Downscaling of Precipitation with Conditional Mixture Models

Julie Carreau
HydroSciences Montpellier, julie.carreau@univ-montp2.fr

Mathieu Vrac
Laboratoire des Sciences du Climat et de l'Environnement

1 Introduction

Statistical downscaling models (SDMs) seek to bridge the gap between large-scale variables simulated from General Circulation Models (GCMs) and small scale variables with high spatial variability such as precipitation. In this paper, we propose to model the distribution of precipitation conditional on large-scale atmospheric information with conditional mixture models (CMMs). CMMs are mixture models whose parameters are computed by a neural network based on large-scale atmospheric predictors. We consider three types of CMMs which differ in the type of continuous densities (Gaussian, Log-Normal or hybrid Pareto) they use as mixture components. We evaluate the three CMMs against the two-component mixture from Williams [3] at downscaling precipitation at three rain gauge stations in the French mediterranean area.

2 Materials and Methods

CMMs combine a discrete component for the "no rain" events and a continuous component for rainfall intensity and can be written as :

$$\phi(y;\psi) = \underbrace{(1-\alpha)\delta(y)}_{\text{no rain}} + \underbrace{\alpha\phi_0(y;\psi_0)}_{\text{rain}>0}, \tag{1}$$

where α is the rain probability, $\delta(\cdot)$ is the Dirac function, $\phi_0(\cdot;\psi_0)$ is the density for rainfall intensity with parameter ψ_0 and $\psi = (\alpha, \psi_0)$. In [3], $\phi_0(\cdot;\psi_0)$ is the Gamma density. We propose to use mixtures instead. We can take into account the dependence of the distribution of precipitation on large-scale atmospheric variables by considering the parameters of the mixture as functions of these variables. A convenient way to implement these functions is by means of a neural network (NN) [1]. The NN parameters are calibrated by minimizing the negative log-likelihood of the conditional mixture over the training set. We selected the hyper-parameters (the number of hidden units and the number of components) via the cross-validation method, see [1]. We evaluate three CMMs which differ in the type of mixture components and compare them with the two-component mixture from Williams [3]. We took Gaussian, Log-Normal or hybrid Pareto ([2]) as mixture components.

The local-scale data are precipitation from three rain gauge stations, Orange, Sète and Le Massegros which are located in the Cévennes-Vivarais, in the French Mediterranean

area. Because of the Mediterranean influence and of the mountainous back country, the Cévennes-Vivarais region is well known for intense rain events, especially in the fall. We have daily rainfall measurements over 46 years (01/01/1959 -12/31/2004) from the *European Climate Assessment & Dataset* (ECA&D). The set of predictors includes the NCEP/NCAR (National Centers for Environmental Prediction/National Center for Atmospheric Research) reanalysis sea level pressure (SLP) fields on a 6 by 6 grid cell regions surrounding the stations. We also include as predictors three date variables representing the year, the month and the week of an observation. Principal component analysis is applied to reduce the dimensionality and remove the redundancy among the predictors. We extract the four principal components in order to keep 90% of the variance of the data.

The 46-year data set is split into a training set of 25 years (01/01/59 - 12/31/83) and a test set of 21 years (01/01/84 - 12/31/04). The training set is first used to select the hyper-parameters with the 5-fold cross-validation method. Then, each model is trained anew on the whole training set with the selected hyper-parameters. The test set serves exclusively for comparison and evaluation of the SDMs.

3 Results

The hybrid Pareto CMM being the most complex model, we first compare the other three SDMs in terms of relative log-likelihood with the hybrid Pareto CMM on the test set. Table 1 shows the relative log-likelihood on the test set along with standard errors for the three competing SDMs on the three rain gauge stations. In bold font are the cases where the hybrid Pareto CMM performed significantly better. We see that the hybrid Pareto CMM outperforms the Gaussian CMM and the Gamma benchmark on all three stations. However, we cannot really distinguish the hybrid Pareto CMM from the Log-Normal CMM based on this criterion.

	Gaussian	Log-Normal	Williams
Orange	**0.02146 (0.003139)**	0.0022512 (0.001910)	**0.02275 (0.002866)**
Sète	**0.01595 (0.003034)**	-0.003530 (0.001647)	**0.01847 (0.002690)**
Le Massegros	**0.01948 (0.006671)**	-0.004606 (0.002121)	**0.02068 (0.003005)**

Table 1: *Relative log-likelihood (std. err.) on the test set between the hybrid Pareto CMM and the other SDMs (Gaussian and Log-Normal CMMs and Williams' model). Positive numbers indicate that the hybrid Pareto CMM performed better. Significant differences are in bold font.*

We randomly generated data for each SDM corresponding to the predictor values on the test set. This was repeated a thousand times. Fig. 1 illustrates the QQ-plots for Orange, on logarithmic scale, between the observations and the simulations for the hybrid Pareto CMM, left panel, and for Williams' model (right panel). Models which are in accordance with the data should be close to the diagonal line. We see that Williams' model is less apt at modelling both the central part (over-estimation) and the upper part (under-estimation) of the distribution. In Fig. 2, we first analyze the seasonal cycles of the rain

probability (left panel) and of the 99% quantile (right panel) of the hybrid Pareto CMM on the Orange test set. We can identify from Fig. 2 two seasonal modes, around March (03) and October (10), which translates into higher probabilities and amounts of rain around these two months, while summer (i.e., around July) presents lower probabilities and amounts of rain. This is globally in agreement with the observations over the test set, showing the same features. In Fig. 3, we finally look at the conditional densities of the hybrid Pareto CMM associated with different atmospheric conditions, that is for different predictors, for the rain event at the Orange station with the highest volume of rain (322 mm in 09/08/2002-09/09/2002) in the test set. The left panel of Fig. 3 shows the central part of the conditional densities while the right panel represents the upper tails in logarithmic scale. Each curve corresponds to a different day which is connected in the legend with the amount of rain observed on that day in chronological order (from top to bottom). From Fig. 3, we see that the conditional density is very responsive to changes in atmospheric conditions and that globally, days with heavy rains correspond to heavy tailed densities and days with no rain to almost flat densities.

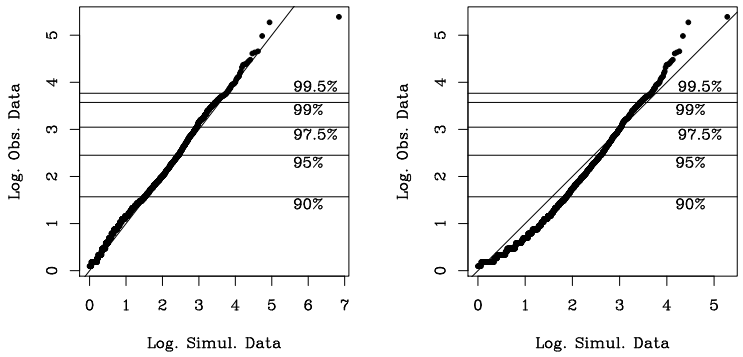


Figure 1: *QQ-plots on logarithmic scale of the simulated precipitation versus observations* $> 1mm$ *on the Orange test set for the hybrid Pareto CMM (left panel), and Williams model (right panel). The horizontal lines are the empirical unconditional quantiles from observations of the test set.*

4 Concluding remarks

To our knowledge, CMMs are used for the first time in a downscaling context and open interesting ways to study the interactions between large- and small-scale climate variables. CMMs extend the two-component mixture proposed initially by Williams [3] which has a discrete component like CMMs to model rainfall occurrence but relies on a single density, the Gamma, for rainfall intensity.

We draw the following conclusions from our analyses on the three stations in the French mediterranean area: 1) CMMs have clear advantages over Williams model in terms of flexibility to represent both the central and the extremal part of rainfall intensity distribution and 2) the choice of component in CMMs depends on the data. In our case, Gaussian

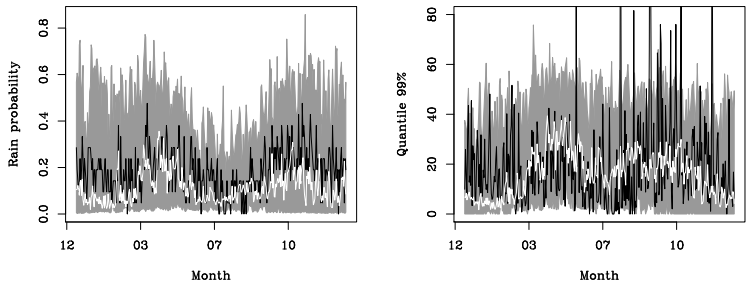


Figure 2: *Daily seasonal cycles of the rain occurrence probability (left panel) and of the 99% quantile (right panel) from the observations (black line) together with an empirical 90% confidence interval (grey band) and median (white line) from the hybrid Pareto CMM for the Orange station test data.*

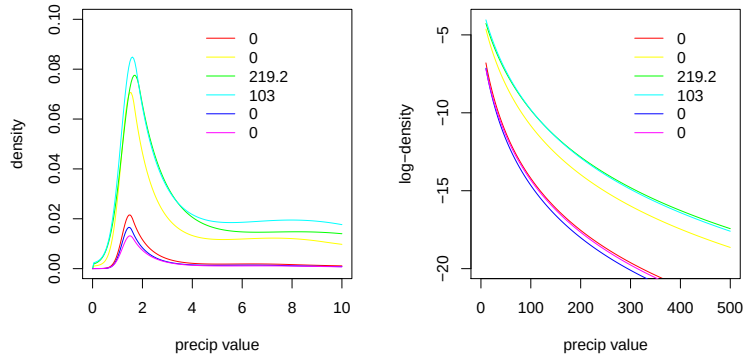


Figure 3: *Conditional densities for the hybrid Pareto CMM day by day for a period with the highest volume of rain in the test Orange data. Each daily density is represented with a different color which is represented in the legend in chronological order, from top to bottom, with the amount of rainfall observed.*

components are not well suited. Log-Normal CMMs offer a good performance and are more straightforward to implement than hybrid Pareto CMMs. However, the assumption of heavy tails of the hybrid Pareto CMM seems more realistic for the precipitation data considered in this work.

References

[1] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford, 1995.

[2] J. Carreau and Y. Bengio. A hybrid Pareto model for asymmetric fat-tailed data: the univariate case. *Extremes*, 12(1):53–76, 2009.

[3] M. P. Williams. Modelling seasonality and trends in daily rainfall data. In *Advances in Neural Information and Processing Systems*, volume 10, pages 985–991, 1998.

A Bayesian Spatio-Temporal framework to improve exposure measurements combining observed and numerical model output

Monica Pirani, John Gulliver, Marta Blangiardo

MRC-HPA Centre for Environment and Health,
Department of Epidemiology and Biostatistics, Imperial College London, UK
E-mail for correspondence: m.pirani@imperial.ac.uk; m.blangiardo@imperial.ac.uk

Abstract: The high resolution Air Dispersion Modelling System (ADSM)-Urban represents an advanced model to simulate the local traffic and non traffic related contribution of PM_{10}. The aim of our study is to provide a Bayesian framework to improve exposure estimates of PM_{10} combining observed data from monitoring sites with ADMS-Urban numerical model output. To illustrate our approach we use PM_{10} daily averaged values for 46 monitoring sites in London, over the period 2002-2003 and output from ADMS-Urban. Different spatio-temporal structures are investigated and compared in performance. We demonstrate that adding covariates on environmental characteristics of sites and meteorological changes over time improve the precision and accuracy of the concentration estimates.

Keywords: Bayesian inference, Particulate matter pollution, Space-Time model, Kriging, Random Walk.

1. Introduction

In the last decade urban air pollution has become a relevant topic of epidemiological and environmental research. The concern over its adverse health effects has led to considerable efforts on the development of numerical model to estimate exposures for these complex mixtures. The high resolution Air Dispersion Modelling System (ADSM)-Urban represents an advanced semi-Gaussian model, widely used to assess and simulate the dispersion into the atmosphere of some important pollutants, such as particulate matter ≤ 10 μm in aerodynamic diameter (PM_{10}), released from industrial, domestic and road traffic sources (Carruthers et al. 2000).
The aim of our study is to provide a Bayesian spatio-temporal framework to improve exposure estimates of PM_{10} combining particulate matter data from monitoring sites with ADMS-Urban model output. Several modelling strategies have been suggested in the Bayesian literature to combine observed data and model output (e.g. Fuentes and Raftery 2005; Sahu et al. 2009; Mc Millan et al. 2010; Berrocal et al. 2010). Our models are framed in a *downscaler* perspective (Berrocal et al. 2010), assuming that PM_{10} is characterised by a spatial and temporal component; we extend this approach incorporating additional relevant spatial or temporal covariates: long-range transport of

PM_{10}, site type, day of the week and temperature. The performance of our modelling approach is assessed using: 1) indexes of model fit and 2) a cross-validation perspective.

2. Materials and Methods

Data description and study area

The dataset consists of PM_{10} daily averaged concentrations ($\mu g/m^3$) that were observed at 46 monitoring sites in London, over the period 2002-2003. The monitoring stations present different environmental conditions, some are in suburban or urban locations (no. 22), and others are located near road (no. 20) or highly busy kerb site (no. 4). The mean distance between the sites is 17813.3 meters (range: 358.4-45297.3 meters). The proportion of missing data is 8.8%, varying across the monitoring sites from 0.7% to 28.4%. The missing values are assumed to be missing at random and being in a Bayesian perspective, they are imputed through the posterior predictive distribution.

The second main source of information is the modelled output for local traffic and non traffic from ADMS-Urban, based on grid cells. It has a limit of 1500 on the number of source road links that can be modelled; monitoring sites were therefore buffered to a distance of 300 metres, and all road sources within that range selected for modelling. Emissions from other sources for each 1 km grid cell were also modelled.

To take into account the contribute of a long-range component of PM_{10}, we included the monitoring station at the rural site of Harwell (~60 Km west of London). Harwell represents a good indicator for long-range transport of air masses: it is surrounded by predominantly agricultural land, and the nearest road is located at 140 metres from the station. In addition, we included in the analysis: the type of site (sub-urban or urban, road and kerb sites), the day of the week (Monday-Friday, Saturday and Sunday or Holiday) and the temperature at the Heathrow meteorological station, measured at 1.25 m above ground level (with linear and quadratic effect).

We performed a preliminary exploratory analysis which showed spatio-temporal variation in the concentration levels of PM_{10}. Figure 1 shows the mean concentration levels: a) by site (quartiles of PM_{10} values distribution) and b) by day for each month (year 2002). The analysis of autocorrelation correlogram of time series (not shown) suggests serial dependencies.

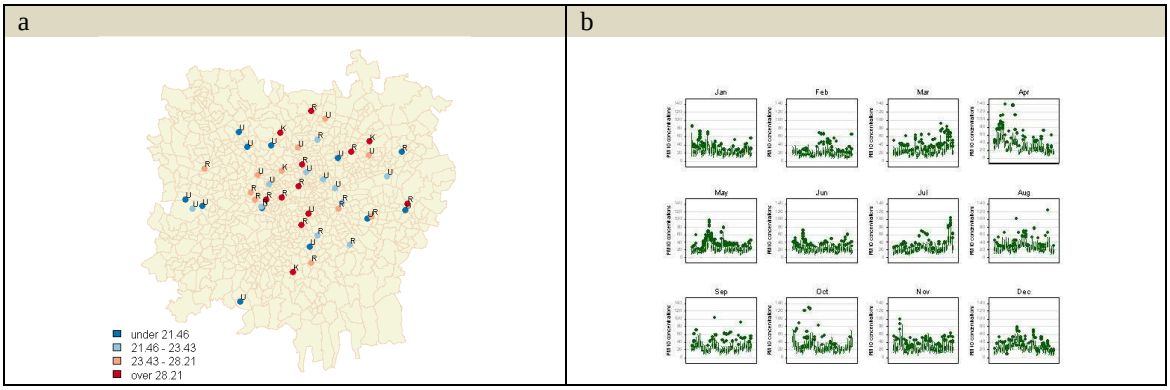


Figure 1: PM_{10} concentrations: a) Plot of mean values by site (U=Urban/Suburban; R=Road; K=Kerb); b) Box plot of daily mean values by month (year 2002)

Bayesian Hierarchical Models

Let Y_{st} denote the response variable (log-transformed PM_{10} data) at location s and time t. The response is modelled as a space-time process defined by $Y_{st} \sim N(\mu_{st}, \sigma_s^2)$.

We consider the following possible models for μ_{st}

Model 1 - $\mu_{st} = \alpha + \beta_{1t} rural_t$

Basic model. Approximately half of the PM_{10} can be considered secondary or natural, being made up of PM formed from gaseous precursors or sea salt, thus this analysis includes only the long-range component (PM_{10} observed at rural site of Harwell) that is assumed to follow a second-order random walk non-stationary in time model.

Model 2 - $\mu_{st} = \alpha + \beta_{1t} rural_t + \beta_{2s} adms_{st}$

Multivariate model that includes, as well as the background component, the output from numerical ADMS-Urban model and its coefficients are assumed to vary spatially through a Bayesian Kriging. We specified a Uniform prior distribution for the correlation decay parameter with range chosen based on prior beliefs about the maximum and minimum correlation at the largest and smallest distances of the PM_{10} values. Prior range for correlation at minimum distance was between 0.10 and 0.99; prior range for correlation at maximum distance was between 0 and 0.30.

Model 3 - $\mu_{st} = \alpha + \beta_{1t} rural_t + \beta_{2s} adms_{st} + \beta_3 type_s + \beta_4 dow_t + \beta_5 temp_t + \beta_6 temp_t^2$

Multivariate model that incorporates spatial and temporal dimension of the data, as well as the spatio-temporal covariates (site type, day of the week, temperature).

We assumed a separate variance for each site σ_s^2 with a moderately informative inverted gamma prior. We adopted vague normal priors for the intercept coefficient α and the regression coefficients $\beta_3, \beta_4, \beta_5, \beta_6$.

To validate our models, we randomly partitioned the monitoring network in four subsets. For each subset, a single subsample is retained as the validation data for testing the model, and the remaining subsamples are used as training data.

The deviance information criterion (DIC; Spiegelhalter et al. 2002), is used to analyse the model fit. In order to compare the performance of the models, we adopted the empirical coverage of 95% credible intervals (95%CI), the average length of 95%CI, the mean square error (MSE), the adjusted R^2 and the mean fractional bias (MFB).

We present the results obtained from one subset; they are consistent for the other subsets.

3. Results

The model comparisons via DIC show large differences among the models: the third one, which considers the spatio-temporal structure as well as the additional covariates, had a smaller DIC (-3506.7) than the first two (DIC respectively equal to 19388.6 and 15574.1). Cross-validation summary statistics are showed in Table 1.

Model	Coverage 95%CI	Average length 95%CI	MSE	Adjusted R^2	MFB
1	95.42	43.15	116.89	0.47	0.10
2	95.66	43.37	106.86	0.47	0.12
3	96.67	32.09	53.81	0.73	0.05

Table 1: Summary statistics for cross-validation prediction

Table 2 presents the posterior distribution of model parameters for Model 3. The effect of the monitoring site type shows that PM_{10} level is significantly higher for road and kerb sites than for suburban/urban sites. Level of PM_{10} are lower on Saturdays (significant) while Sunday or Holidays are not significantly different from weekdays. High temperatures are associated with high concentration of PM_{10}. Finally, the relationship between observed values and modelled output from ADMS-Urban shows spatial variation (Figure 2).
The posterior median of daily temporal effect (parameter β_{1t}) associated with long-range component (not shown) presents a range of values from -1.36 to 1.39 (95%CI).

Parameters	Median	2.5%	97.5%
α	2.787	2.723	2.837
β_3 (Road site)	0.150	0.143	0.158
β_3 (Kerb site)	0.220	0.205	0.233
β_4 (Saturday)	-0.219	-0.271	-0.172
β_4 (Sunday or Holiday)	0.070	-0.006	0.147
β_5 (Temperature)	0.122	0.112	0.147
β_6 (Temperature^2)	0.021	0.019	0.026

Table 2: Posterior distribution of model parameters (on log-scale)

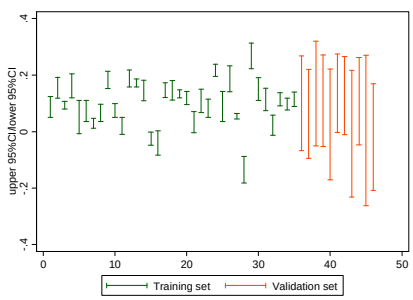


Figure 2: Posterior distribution of β_2 parameter (on log-scale)

4. Concluding remarks

Our Bayesian approach provides a natural way to combine data from different sources taking into account their uncertainties. We found that adding "spatial" covariates (e.g. site type) and "temporal" ones (day of the week, temperature) increases the precision and accuracy of the estimated values of PM_{10}.

References

Berrocal V. J., Gelfand A. E., Holland D. M. (2009) A Spatio-Temporal Downscaler for Output from Numerical Models, *Journal of Agricultural, Biological, and Environmental Statistics*, 15, 176-197.

Carruthers D.J., Edmunds H.A., Lester A.E., McHugh C.A., Singles R.J. (2000) Use and validation of ADMS-Urban in contrasting urban and industrial locations, *International Journal of Environment and Pollution*, 14 (1–6), 364–374.

Fuentes M., Raftery, A. E. (2005) Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models, *Biometrics*, 61 36–45.

McMillan N., Holland D. M., Morara M., Feng J. (2010) Combining Numerical Model Output and Particulate Data Using Bayesian Space-Time Modeling, *Environmetrics*, 21, 48-65.

Sahu S. K., Yip S., Holland D. M. (2009) Improved space–time forecasting of next day ozone concentrations in the eastern US, *Atmospheric Environment*, 43, 494-501.

Spiegelhalter D., Best N., Carlin B., van der Linde A. (2002) Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society, Series B*, 64, 583-639.

A Spatio-temporal model for cancer incidence data with zero-inflation [1]

Monica Musio
Department of Mathematics, University of Cagliari, Italy, mmusio@unica.it

Erik A. Sauleau
Faculty of Medicine, University of Strasbourg, France

Abstract: In this work we consider a joint space-time model for cancer incidence, using data on prostate cancer collected between 1988 and 2005 in a specific area of France. Our aim is to take into account possible non linear effects of some covariates and zero-inflation due to data aggregation for Poisson regression. We assume that counts of cancer cases follow zero-inflated Poisson distribution, where the probability of zero inflation is a monotonic function of the mean. The purpose of our analysis is to check whether the French prostate screening programme, which begins in 1994, results in a spatial or a spatial-temporal change of the pattern of the disease.

Keywords: Spatio-temporal model, cancer incidence data, zero inflation

1 Introduction

Cancer registries represent epidemiological instruments which are aimed at providing population based cancer incidence and mortality summaries. Usually the data are stratified by age group, year and geographical unit of residence. As the counts of cancer cases are distributed according to these variables, the dataset exhibits a proportion of zeros higher than would be expected under the Poisson distribution. The problem is also known as zero-inflation (Lachenbruch, 2002) and is common in ecological studies. We make the assumption, justified by the nature of the data analyzed, that the probability of zero inflation depends on the set of stratified variables. In this work we analyse data on prostate cancer incidence collected between 1988 and 2005 in the North-East of France. We present an approach to analyze the space-time evolution of the disease taking into account also possible non linear effects of other covariates (such as age) and the zero inflation due to extra Poisson variation. Prostate is a type of cancer which usually does not have a spatial distribution. Here we are interested in the space-time evolution of the disease to investigate if the prostate screening programme started progressively in the region since 1994 has a direct implication on the space or space-time evolution of the cancer.

[1]The second author was partially supported by *Visiting Professor program* of "Regione Autonoma della Sardegna"

2 Materials and Methods

Our data consists of all cases of prostate cancer (C61.- in the ICD-10 classification) diagnosed between the 1st January 1988 and 31st December 2005, in the region of Haut-Rhin in France. The total number of cases is 6878. The distribution of the number of cases aggregated over age groups (9 categories), across the 26 geographical units, each year has mean of 14.2 cases while the median is 10. Due to covariates, the data set counts were spread over 4374 cells with 1935 zeros (44% of the cells are equal to zero). Our objective is to detect effects of time, space, age and age-time interaction on the number of new prostate cancer cases, taking into account an high proportion of zero counts. We thus build different zero-inflated models and compare them using marginal likelihood.

Zero-inflated Poisson data are often analyzed via a mixture model specifying that the response variable, Y, comes from a mixture of 0 with probability ω and a regular Poisson component of mean λ with probability $1-\omega$ (Lambert, 1992).

Covariates may then enter into the model through the mean λ and/or through the probability ω. Here we consider a zero-inflated generalized additive model (Chiogna and Gaetan, 2007), where the mean of the regular component and the probability of zero-inflation are each modeled as a function of some nonparametric smooth predictors. As usual we assume that the mean of the Poisson distribution λ is equal to $\mathrm{E}(\mu)$ where E indicates expected number of cases under direct standardization and μ is the relative risk. For the log risk we consider the following linear predictor:

$$\log(\mu_{atr}) = \eta_{atr} = f_1(age_a) + f_2(year_t) + f_3(age_a, year_t) + f_4(east_r, north_r) \quad (1)$$

$a \in \{1,\ldots,9\}$, $t \in \{1,\ldots,18\}$, $r \in \{1,\ldots,26\}$, $f_1(\cdot)$, $f_2(\cdot)$ are smooth functions of the covariates age and year modeled using cubic regression splines, $f_4(east_r, north_r)$ is a thin plate regression spline, while, for modelling the smoothed age-time interaction, we use tensor products allowing smoothness parameter selection to be independent of the different scale of the covariates (for more details see (Wood, 2006)). We make the assumption that the probability of zero inflation is a linear function of the covariates. We are in the framework of constrained zero-inflated generalized additive model (COZIGAM) ((Liu and Chan, 2010)). In particular we consider the following two specifications:

1. Model 1: the dependence is constrained in such a way that the probability of zero inflation is linearly related to linear predictor. We have:

$$logit(\omega_{atr}) = \alpha + \delta\eta_{atr};$$

2. Model 2: the proportional constraint can be generalized by assuming that the proportionality constant is specific to each additive component, specifically:

$$logit(\omega_{atr}) = \beta + \delta_1 f_1(age_a) + \delta_2 f_2(year_t) + \delta_3 f_3(age_a, year_t) + \delta_4 f_4(east_r, north_r).$$

In both model the linear predictor is specified as in equation (1).

Because there is no closed form for the marginal likelihood, Laplace method is used to approximately compute the likelihood (Liu and Chan, 2010). The analyses have been performed using the R package `COZIGAM` (Liu and Chan, 2010), relying on `mgcv` package (Wood, 2001).

3 Results

According to the marginal likelihood, the best model is Model 2. In Table 1 are reported the values of the significant proportionality coefficients estimates for the best model, which provides strong evidence of a significant relationship between these smooth components in the mean of the non-zero-inflated distribution and in the zero-inflation probability, on their link scales. These values emphasize the main role that age plays on the zero-inflation, compared with the effect of time.

Figure 3 displays the smooth function estimates of Model 2. We can see that:

- The estimate of the time effect shows an increase of incidence up to 1995 then a strong decrease up to 2001 then an increase.
- the combined effect of age and time is quite relevant, in particular a progressive decrease in the age for the maximum incidence along time is evident.
- The estimated spatial effect is slightly significant. Except some boundary effects, there is a little peak of incidence in the north of the region (where a city of around 70,000 inhabitants is) and again a peak on the south-east part, difficult to separate from the boundary effect. Adding the spatio-temporal interaction in the model mod4 yields a non-significant effect.

	Covariate	estimate	standard error	pvalue
β		3.9939	0.85714	$p < 0.00001$
δ_1	s(age)	0.8859	0.04692	$p < 0.00001$
δ_2	s(year)	2.595	1.150	$p = 0.024$
δ_3	s(age, year)	1.237	0.29	$p < 0.00001$

Table 1: Significant coefficient estimates of the constrained generalized additive model

4 Concluding remarks

Zero-inflated generalized additive model provides a method for modeling incidence data by taking simultaneously into account possible non linear effects of continuous covariates and the spatio-temporal evolution of the disease. The number of extra zeros seems in particular linked to the age group. The aim of such study is to check

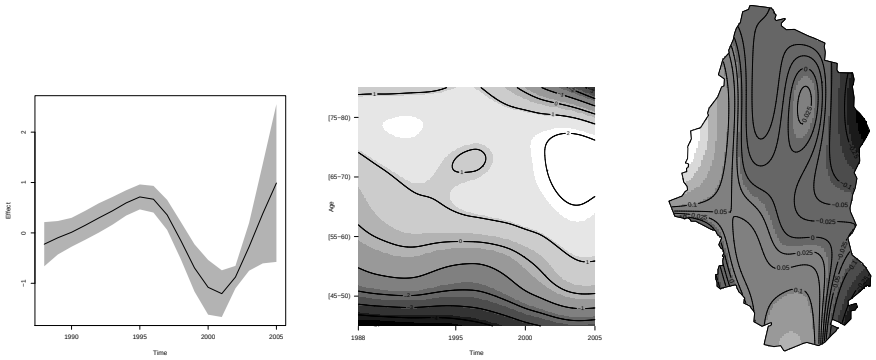


Figure 1: Effect of time, joint effect of age and time and spatial effect estimated for Model 2

whether the spatial pattern of incidences changes over time. The main finding is that there is a strong temporal effect, while the spatial effect is not very strong (not quite significant) and the spatial effect does not change over time (the space-time interaction was not significant). If we link the aspect of the main temporal effect with the development of the screening, it seems that the effect on the prostate cancer incidence is relevant since 1998 whereas the beginning of the organized screening campaign is 1994. This difference is probably due to a certain time for the screening programm to be fully efficient in the population.

References

Chiogna, M. and Gaetan, C. (2007). Semiparametric zero-inflated poisson models with application to animal abundance studies. *Environmetrics*, 18:303–314.

Lachenbruch, P. (2002). Analysis of data with excess zeros. *Statistical Methods in Medical Research*, 11:297–302.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34:1–14.

Liu, H. and Chan, K. (2010). Introducing COZIGAM: an R package for unconstrained and constrained zero-inflated generalized additive model analysis. *Journal of Statistical Software*, 35(11):1–26.

Wood, S. (2001). mgcv: GAMs and generalized ridge regression for R. *R News*, 1(2):20–5.

Wood, S. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62:1025–1036.

Generalized Estimating Equations for Zero-Inflated Spatial Count Data[1]

Anthea Monod
Department of Mathematics, Swiss Federal Institute of Technology (EPFL),
Station 8, CH-1015 Lausanne, Switzerland; anthea.monod@epfl.ch
Under the supervision of Prof. Stephan Morgenthaler

Abstract: We consolidate the zero-inflated Poisson model for count data with excess zeros (Lambert, 1992) and the two-component model approach for serial correlation among repeated observations (Dobbie and Welsh, 2001) for spatial count data. This concurrently addresses the problem of overdispersion and distinguishes zeros that arise due to random sampling from those that arise due to inherent characteristics of the data. We give a general quasi-likelihood and derive corresponding score equations for the zero-inflated Poisson generalized linear model. To introduce dependence, a spatial-temporal correlation structure comprising forms for fixed time, fixed location, and neighbor interactions is required; construction using techniques from the theory of Markov point processes is investigated.

Keywords: Generalized estimating equation (GEE), spatial count data, zero-inflated counts, zero-inflated Poisson model, nearest-neighbor marked Markov point processes, Dirichlet tessellation.

1 Introduction

Let y_{it} denote the number of occurrences of an event observed at $t = 1, \ldots, T_i$ time points for each subject $i = 1, \ldots, n$, and let $\mathbf{x}_{it} \in \mathbf{R}^q$ be a vector of measured covariates. Such data is often modeled through a generalized linear model to provide greater flexibility, specifying a form for the expectation, $E[Y_{it}] = \lambda_{it} = g^{-1}\big(\mathbf{x}'_{it}\boldsymbol{\beta}\big)$, with $\boldsymbol{\beta}$ a $q \times 1$ vector of unknown parameters, and the link function $g(\cdot)$ commonly taken to be the log function. Under a Poisson distribution, the variance is equal to the mean, $\mathrm{Var}(Y_{it}) = \lambda_{it} = E[Y_{it}]$, which in practice may be too restrictive; often the data exhibit $E[Y_{it}] = \mathrm{Var}(Y_{it})$, known as overdispersion.

Lambert (1992) has presented zero-inflated Poisson (ZIP) regression, giving rise to a new class of regression models for count data with an abundance of zero observations. In a ZIP model, the non-negative integer response Y is assumed to be distributed as a mixture of a Poisson distribution with parameter λ_{it}, and a distribution with point mass of one at the value zero, with mixing probability α_{it}; the non-zeros and a portion of the zeros are modeled by the usual Poisson probability.

[1]Research supported in part by the Swiss National Science Foundation, Grant No. FN 200021-116146

Dobbie and Welsh (2001) adapt the generalized estimating equations approach of Liang and Zeger (1986) to zero-inflated spatial count data, addressing dependence by incorporating a correlation matrix. They model the abundance of zeros via a two-component approach: the zeros are modeled separately from the non-zeros; first, absence versus presence (zero versus non-zero) is described by a logistic model, and then conditional on presence, the non-zero counts are described by a truncated Poisson distribution.

We work in the context of a Poisson generalized linear model, consolidating the two aforementioned approaches to construct generalized estimating equations for the zero-inflated Poisson generalized linear model comprising spatial-temporal dependence. Attributing some of the zeros to the Poisson distribution avoids conditioning on the responses, and provides a more intuitive approach to occurrence of zeros in the data. The data of interest are weekly counts of Noisy Friarbirds (*Philemon corniculatus*) recorded by observers for the Canberra Garden Bird Survey: attributing a probability weight of zero observations to a point mass distribution and its complement to a Poisson distribution allows for the distinction between zero counts arising due to an inherent characteristics that may induce zero observations (*e.g.* inadequacy of the region where measurements were taken for the survival or reproduction of Noisy Friarbirds), and zero counts arising at random. In considering dependence, the theory of nearest-neighbor Markov point processes proves to be useful in constructing covariance forms for the zero-inflated spatial data.

In this paper, we detail the theoretical results behind the work to be presented at the 2011 European Regional Conference of The International Environmetrics Society (TIES), "Spatial Data Methods for Environmental and Ecological Processes – 2nd Edition".

2 Methodology

We implement the zero-inflated Poisson model of Lambert (1992) to address overdispersion, and obtain a likelihood and score equations, which, following Dobbie and Welsh (2001), turn out to be generalized estimating equations in the style of Liang and Zeger (1986); we incorporate dependence into the model following Diggle *et al.* (2009). In constructing a space-time dependence structure, we focus on the neighbor interaction component and outline the theory of Markov point processes relevant to this aspect.

The Zero-Inflated Poisson Generalized Linear Model. A non-negative, integer-valued random variable describing a discrete number of occurrences for a cross-sectional unit i at time period t is said to follow a *zero-inflated Poisson distribution* with parameter $\lambda_{it} \in (0, \infty)$ and mixing probability $\alpha_{it} \in (0, 1)$ if

$$Y_{it} \sim \begin{cases} 0 & \text{with probability } \alpha_{it}, \\ \text{Poisson}(\lambda_{it}) & \text{with probability } (1 - \alpha_{it}). \end{cases} \tag{1}$$

The parameters $\lambda_{it}, \alpha_{it}$ are allowed to depend on auxiliary covariate information, for simplicity we assume the same auxiliary information. It follows that $E[Y_{it}] = (1-\alpha_{it})\lambda_{it}$ and $\text{Var}(Y_{it}) = (1-\alpha_{it})\lambda_{it}(1+\alpha_{it}\lambda_{it})$ and indeed $E[Y_{it}] < \text{Var}(Y_{it})$.

Under this model, the observations are generated by

$$\text{Prob}(Y_{it} = y_{it}|\mathbf{x}_{it}) = \alpha_{it}\mathbb{1}(y_{it}=0) + (1-\alpha_{it})\frac{\exp\left\{y_{it}\mathbf{x}_{it}'\boldsymbol{\beta} - e^{\mathbf{x}_{it}'\boldsymbol{\beta}}\right\}}{y_{it}!}, \tag{2}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function; the probability of observing a zero is $\text{Prob}(Y_{it}=0|\mathbf{x}_{it}) = \alpha_{it} + (1-\alpha_{it})\exp\left\{-e^{\mathbf{x}_{it}'\boldsymbol{\beta}}\right\}$.

Likelihood and Score Equations. The log-likelihood for the zero-inflated Poisson model is $\ell(\alpha_{it}, \boldsymbol{\beta}) = \sum_{i,t:y_{it}=0} \log\left(\alpha_{it} + (1-\alpha_{it})e^{-e^{\mathbf{x}_{it}'\boldsymbol{\beta}}}\right) + \sum_{i,t:y_{it}>0}\left(y_{it}\mathbf{x}_{it}'\boldsymbol{\beta} - e^{\mathbf{x}_{it}'\boldsymbol{\beta}} - \log y_{it}!\right)$, which gives the following score equation with regard to $\boldsymbol{\beta}$:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\ell(\alpha_{it},\boldsymbol{\beta}) = \sum_{i,t:y_{it}=0}(y_{it}-\lambda_{it})\frac{\text{Prob}(Y_{it}=0)-\alpha_{it}}{\text{Prob}(Y_{it}=0)} + \sum_{i,t:y_{it}>0}(y_{it}-\lambda_{it})\mathbf{x}_{it} = 0. \tag{3}$$

Modeling the mixing probability α_{it} as any differentiable function of another parameter γ, $\alpha_{it} = \alpha_{it}(\gamma)$, the score equation for the ZIP model with regard to γ is

$$\frac{\partial}{\partial\gamma}\ell(\alpha_{it},\boldsymbol{\beta}) = \sum_{i,t}\frac{\text{Prob}(Y_{it}>0)}{\text{Prob}(Y_{it}=0)}\frac{\partial\alpha_{it}}{\partial\gamma}\frac{1}{1-\alpha_{it}} = 0. \tag{4}$$

Note that ratio of probabilities in this latter equation provides an intuitive odds-ratio interpretation of the weighting between the two probability components.

Introducing Dependence. Following Dobbie and Welsh (2001) and Diggle *et al.* (2009) in the setting of marginal models, we introduce dependence by extending the score equations (3) and (4) to comprise a $2T_i \times 2T_i$ spatial variance-covariance matrix. Diggle *et al.* (2009) show that for marginal models under appropriate parameterizations, the score equations assume a form of a generalized estimating equation (Liang and Zeger, 1986), whose solution gives a consistent estimator:

$$\left(\frac{\partial\mu}{\partial\boldsymbol{\beta}}\right)'\text{Var}(Y)^{-1}(Y-\mu) = 0. \tag{5}$$

In the spatial-temporal setting, the covariance requires structures for fixed time, fixed location, and neighboring interactions. Models for the former cases are readily available in time series analysis and spatial statistics literature. In our application to Noisy Friarbird counts, the latter case is of particular interest, since, depending on the region partition, observations in one region is likely to influence that in nearby regions: vicinities of unsuitable habitat regions may also be less suitable, thus influencing a low-valued observation. This motivates the use of techniques

of nearest-neighbor Markov point processes and random tessellations to address neighbor interaction as well as region partitioning.

Models for observations generated by marked Markov point processes can be augmented to allow interactions to depend on the realization of the process by generalizing the spatial Markov property, as shown by Baddeley and Møller (1989). Moreover, the spatial interaction in a marked Markov point process can be analyzed conditional on the positions of the points, since the conditional distribution of the marks given the point configuration is a Gibbs process on the finite graph defined by the points. Dirichlet tessellation is shown to satisfy nearest-neighbor conditions in the construction of such processes where each point interacts with its neighbors, notably that of the invariance of connectivity between any two points under the addition of a new point, unless it is a neighbor of both points.

3 Concluding Remarks

The spatial-temporal zero-inflated Poisson generalized linear model addresses overdispersion present in space-time data comprising excess zeros, while providing greater flexibility in the modeling and interpretation of zeros due to random sampling and those due to characteristics of the data, and may be extended to incorporate spatial-temporal dependence. The nature of such data motivates the consideration of neighboring interactions when constructing forms for dependence, which then inspires the use of techniques of nearest-neighbor Markov point processes, allowing for the generation of spatial points with interaction that is conditional on their positions. This two-fold approach to the challenges of zero-inflated, correlated spatial-temporal data will indeed prove to be applicable in various ecological and biological contexts, and useful in general applications in diverse fields of science.

References

Baddeley A., Møller J. (1989) Nearest-Neighbour Markov Point Processes and Random Sets, *International Statistical Review*, 57(2), 89–121.

Diggle P. J., Heagerty P., Liang K.-Y., Zeger S. (2009) *Analysis of Longitudinal Data* 2nd ed., Oxford University Press, Oxford.

Dobbie M. J., Welsh A. H. (2001) Modelling Correlated Zero-Inflated Count Data, *Aust. N. Z. Stat.*, 43(4), 431–444.

Lambert D. (1992) Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing, *Technometrics*, 34(1), 1–14.

Liang K.-Y., Zeger S. (1986) Longitudinal Data Analysis Using Generalized Linear models, *Biometrika*, 73(1), 13–22.

Poisson M-Quantile Geographically Weighted Regression on Disease mapping

Ray Chambers
Centre for Statistical and Survey Methodology, University of Wollongong,
New South Wales 2522, Australia email: ray@uow.edu.au

Emanuela Dreassi
Dipartimento di Statistica "G. Parenti", Università degli Studi di Firenze,
Viale Morgagni, 59 - 50134 Florence, Italy email: dreassi@ds.unifi.it

Nicola Salvati
Dipartimento di Statistica e Matematica Applicata all'Economia,
Università di Pisa, Via Ridolfi, 10 - 56124 Pisa, Italy email: salvati@ec.unipi.it

Abstract: A new approach to ecological analysis on disease mapping is introduced: a semi-parametric approach based on M-quantile models. We define a Poisson M-Quantile spatially structured model. The proposed approach is easily made robust against outlying data values for covariates. Robust ecological disease mapping is desirable since covariates at area level usually present measure-type error. We consider a spatial structure in the model in order to extend the M-quantile approach to account for spatial correlation between areas using Geographically Weighted Regression (GWR). Differences between M-quantile and usual random effects models are discussed and the alternative approaches are compared using the Scottish Lip cancer example.

Keywords: disease mapping, ecological analysis, M-quantile regression, Robust models, spatial correlation, Poisson regression, geographically weighted regression

1 Introduction

Disease mapping involves the analysis of disease incidence or mortality data often available as aggregate counts over a geographical region subdivided for administrative purposes. Such aggregate data are often relatively easy to obtain from government sources. More difficult is to obtain the measures, at aggregated level, on explanatory covariates that could be considered as known or putative risk factors.

Ecological regression on disease mapping mainly focuses on the estimation of risk in administrative regions and the analysis of the association between risk factors and disease. In ecological analysis related to disease mapping, data usually exhibit overdispersion. The latter is usually considered in the model by way of random effects introduced on the model. Clayton and Kaldor (1987) proposed the use of a Poisson-

gamma model for relative risks using an Empirical Bayesian approach (referred to as EB below). This model was generalized by Besag *et al.* (1991) into a fully Bayesian setting using a Hierarchical Bayesian model with a spatial structure (referred to as BYM below). So, ecological disease mapping typically rely on regression models that use both covariates and random effects to explain variation between areas. These models depend on strong distributional assumptions and require a formal specification of the random part of the model. Moreover, they do not easily allow for outlier-robust inference due to covariates at areal level that could be measure-type error prone.

In this paper, we describe a new approach to ecological disease mapping: Poisson M-Quantile regression (referred to as PMQ below). Raughly speaking, the idea is to model quantiles like parameters of the conditional distribution of the target variable given the covariates. Unlike usual random effects models, M-quantile models do not depend on strong distributional assumptions and are robust to the presence of outliers due to measure-type error on covariates. We introduce easily a spatial structure extending the M-quantile approach to account for such spatial correlation between areas by way of appropriate weights at the estimation step (see Salvati *et al.*, 2011). The used approach to incorporate such spatial information is Geographically Weighted Regression: the relationship between the outcome variable and the covariates is characterised by local rather than global parameters, where local is defined spatially. Differences between Poisson M-quantile and traditional random effects models are discussed and compared using the Scottish Lip cancer example.

2 Poisson M-Quantile regression

We define an extension of linear M-quantile regression to count data. M-quantile regression (Breckling and Chambers, 1988) is a "quantile-like" generalization of regression based on the influence function (M-regression). The M-quantile of order q, $q \in (0,1)$, of a random variable Y with continuous distribution function $F(\cdot)$ is the value Q_q that satisfies

$$E\left[\psi_q\left(\frac{Y-Q_q}{\sigma_q}\right)\right]=0$$

where σ_q is a suitable measure of the scale of the random variable $Y-Q_q$, $\psi_q(\epsilon)=2\psi(\epsilon)\left[qI(\epsilon>0)+(1-q)I(\epsilon\leq 0)\right]$ and ψ is an appropriately chosen influence function: the Huber "small c" second proposal specification with $c=1.345$, $\psi(\epsilon)=\epsilon I(-c\leq\epsilon<c)+c\ sgn(\epsilon)I(|\epsilon|>c)$.

Breckling and Chambers (1988) define a linear M-quantile regression model as one where the M-quantile $Q_q(X;\Psi)$ of the conditional distribution of Y given the matrix of p auxiliary variables X corresponding to an influence function ψ satisfies

$$Q_q(X;\psi)=X\beta_{q\psi}$$

There is no agreed definition of an M-quantile regression function when Y is rates parameterized Poisson. The most appealing, of course, is using a log-linear specification

$$Q_q(X;\psi) = t\exp(\gamma_{q\psi})$$

where $\gamma_{q\psi} = X\beta_{q\psi}$ is the linear predictor and t the offset term (expected cases of death). Cantoni and Ronchetti (2001) obtained a robust version of the estimating equations for generalized linear models. We consider the extensions of this to the M-quantile geographically weighted regression case (referred to as PMQGWR below) following Salvati *et al.* (2011).

3 Scottish Lip cancer Example

Clayton and Kaldor (1987) and many others (i.e. Wakefield, 2007) analyzed observed and expected numbers of lip cancer cases in the 56 administrative areas of Scotland. Data were available on the percentage of the work force in each county employed in agriculture, fishing or forestry. This covariate have been chosen because all three occupations involve outdoor work, exposure to sunlight, the principal known risk factor for lip cancer. In the present paper, we analyse this data using EB, BYM using a convolution prior (exchangeable and spatially structured random terms), PMQ and PMQGWR models. Figure 1 shows estimates of relative risk for considered models. Results are similar. Poisson M-quantile models, seems smoother less than random effects models. For PMQGWR sensitivity analysis to bandwidth choice has to be considered.

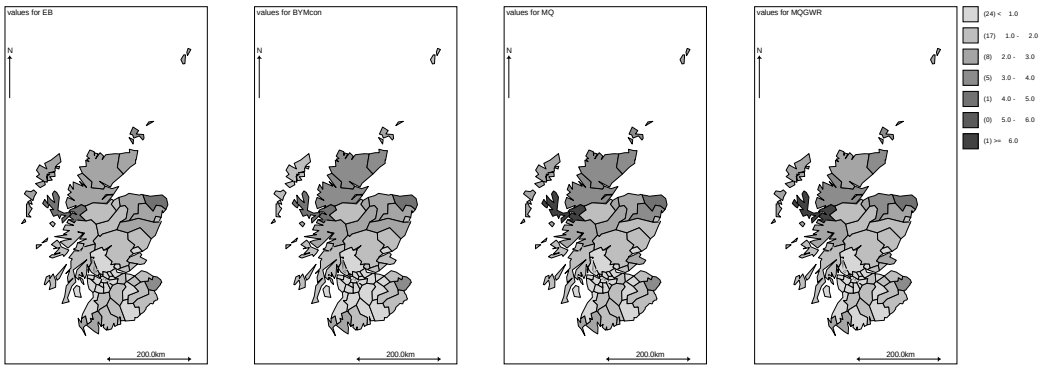


Figure 1: Relative risks estimates using different models: EB, BYM gaussian convolution, PMQ and PMQGWR

4 Conclusion

In this paper, M-quantile models for ecological analysis on disease mapping are introduced and investigated. In particular, we specify an M-quantile GWR model that is a local model for the M-quantiles of the conditional distribution of the outcome variable given the covariates. This model is then used to define a bias-robust predictor of the small area characteristic of interest that also accounts for spatial association in the data. These models offer a natural way of modeling between area association and variability without imposing prior assumptions about the source of this variability. In particular, with M-quantile models there is no need to explicitly specify the random components of the model; rather, inter-area differences are captured via area-specific M-quantile coefficients. As a consequence, the M-quantile approach reduces the need for parametric assumptions. In addition, estimation and outlier robust inference under these models is straightforward. The proposed approach appears to be suitable for estimating a wide range of parameters and our simulation results show that it is a reasonable alternative to mixed effects models for ecological analysis on disease mapping.

References

Besag J., York J.C., Mollie A. (1991) Bayesian image restoration, with application in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics*, 43, 1-59.

Breckling J., Chambers R. (1988) M-quantiles, *Biometrika*, 75, 761-771.

Cantoni E., Ronchetti E. (2001) Robust Inference for Generalized Linear Models, *Journal of the American Statistical Association*, 96, 1022-1030.

Clayton D., Kaldor J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics*, 43, 671-681.

Lawson A.B., Browne W.J., Vidal Rodeiro C.L. (2003) *Disease Mapping with WinBUGS and MLwiN*, Wiley, London.

Salvati N., Tzavidis N., Pratesi M., Chambers R. (2011) Small area estimation via M-quantile Geographically Weighted Regression, *Test*, forthcoming (DOI 10.1007/s11749-010-0231-1).

Wakefield J. (2007) Disease mapping and spatial regression with count data, *Biostatistics*, 8, 158-183.

A software for optimal information based downsizing/upsizing of existing monitoring networks.

Emanuele Barca, Giuseppe Passarella, Michele Vurro
Water Research Institute - National Research Council, CNR-IRSA
V.le De Blasio, 5 - 70132 Bari (Italy)
emanuele.barca@ba.irsa.cnr.it

Alberto Morea
Dept. of Physics - University of Bari
Via Amendola, 173 - 70126 Bari (Italy)

Abstract: Using reliable stochastic or deterministic methods, it is possible to rearrange an existing network by eliminating, adding or moving monitoring locations producing the optimal arrangement among any possible. In this paper, some spatial optimization methods have been selected as more effective among those reported in literature and implemented into a software M-Sanos able to carry out a complete redesign of an existing monitoring network. Both stochastic and deterministic methods have been embedded in the software with the option of choosing, case by case, the most suitable with regard to the available information. Finally, an application to the existing regional groundwater level monitoring network of the aquifer of Tavoliere located in Apulia (south Italy) is presented.

Keywords: Environmental monitoring, Spatial simulated annealing, Kriging.

1. Introduction

With the growth of public environmental awareness and the contemporary improvement in national and EU legislation regarding the environment, monitoring has assumed great importance in the frame of all those managerial activities related to environmental protection and safeguarding. The recent technical and scientific literature has produced a huge amount of papers related to the Optimal Monitoring Network Redesign (OMNR) (Barca et al., 2008; Wu, 2004). Typical OMNR problems consist in adding, removing or moving one or more measurement point in the monitoring network. Scientific literature often refers to these cases as upsizing, downsizing and relocation. In general, the OMNR is an optimization problem solvable through the quantitative formulation of one or more objective functions (OF), whose minimization can be achieved iteratively through various network configurations that meet specific conditions of theoretical and practical nature. The choice of the OF strongly depends on the goals and the information available. Among the iterative optimization methods, one of the most cited in the literature is the so-called Spatial Simulated Annealing (SSA) (Kirkpatrick et al., 1983; Van Groenigen et al., 2000). Many of the methods developed for OMNR require a huge computational effort, consequently, some authors developed software able to perform this task which, however, generally deals only with one of the possible aspects of OMNR (Hu and Wang, 2010; Naoum and Tsanis, 2004; Jimênez et al., 2005; Van Groenigen and Stein, 1998; Passarella et al., 2003).

This paper presents a software developed in MATLAB able to solve any OMNR problem. It allows one to use several approaches (deterministic, stochastic, mixed), techniques (SSA, Greedy deletion) and OF (kriging variance estimation, geometric parameters). A case study based on the downsizing of the groundwater level monitoring network of the Apulia Region located in the aquifer of Tavoliere (Southern Italy) is presented. Three piezometric stations have been removed from the existing monitoring network, made of 30 measurement stations.

2. Materials and Methods

The proposed software is fundamentally made by a Network Downsizing Module and a Network Upsizing Module, which allows one to solve 3 different OMNR related problems: (i) removing points from an existing monitoring network; (ii) adding new points to the monitoring network; (iii) moving points from the existing location to another one as a combination of (i) and (ii). The proposed software provides suitable techniques able to produce reliable optimal solutions to different OMNR problems once the goals have been focused and the available information has been evaluated. The backbone of the software is the Spatial Simulated Annealing (SSA).
Other three modules complete the software architecture: an Input Module, an Output Module and an optional Variography Module (Optional). The input module has been designed in order to support the user in this phase which is strongly dependent from the problem, the goals and the available data. An optional variography module has been added to the software capable of performing a best fit of a model to the experimental variogram. Once the input phase has been completed the software starts running, showing, real time, the evolution of the current transitory optimal configurations. The output of the software consists in a list of the coordinates of the redesigned monitoring network together with some statistics and plots representative of the convergence rate of the method. The software has been named M-SANOS (MATLAB SANOS) in order to honour the well known software SANOS (Spatial ANnealing for Optimal Sampling) proposed by Van Groenigen and Stein (1998) which is the first approach to OMNR based on SSA. Starting from SANOS, new options have been implemented in M-SANOS, as the downsizing module, new OFs and heuristics. Several study cases have been implemented in order to test the software reliability and efficiency. As an example, a case study referred to the downsizing of the groundwater levels monitoring network, consisting of 30 piezometers, and located in the aquifer of Tavoliere in the Apulia Region (Italy) is presented. The study area extends over 1275 km^2 and it corresponds to the largest alluvial plain of southern Italy (fig.1). The simulation concerned the elimination of three wells from the original configuration.

3. Results

The Mean KEV (kriging estimation variance) has been used as OF for the case study simulation to reach the goal of increasing the accuracy for kriging estimations to be carried out at unsampled points over the monitored area. Figure 1 shows the study area and the starting monitoring network. A gaussian variogram model has been fitted to the experimental data. After about 1300 iterations, the method converged to the optimal configuration characterized by a correspondent value of the OF of 0.433. Figure 2 shows the resulting configuration; the three empty dots are those removed by the

optimization method, while Figure 3 shows the behaviour of the correspondent value of the OF (*fitness value*) vs. the iterations and the final configuration.

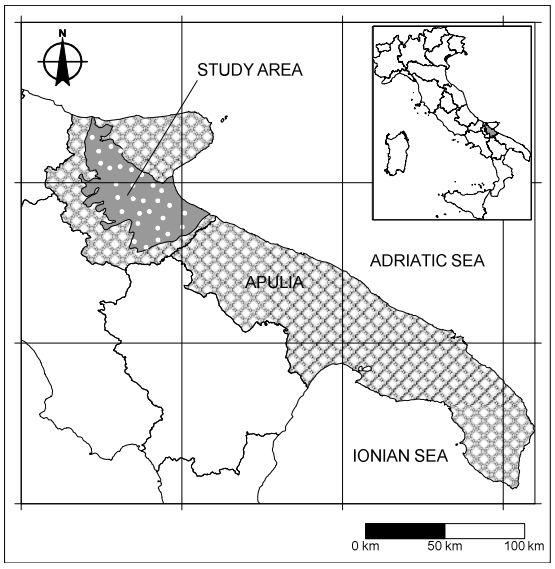


Figure 1: Study area.

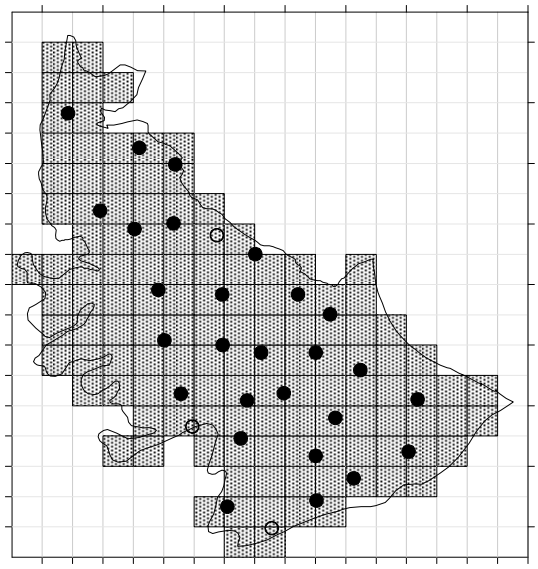


Figure 2: Results of the simulation: greyed background = simulation grid; empty dots = removed monitoring points.

In order to evaluate the effectiveness of the method applied, the optimal network configurations has been verified through complete enumeration. In practice, all the possible $\binom{30}{3} = 4060$ configurations have been generated and the minimum value of the OF has been evaluated. This value corresponds exactly to the *fitness value* of the optimal configuration resulting from the simulation. Figure 4 shows the values of the OF (mean KEV) for all the possible configurations of the monitoring network in descending order. It confirms that the minimum value, corresponding the global optimum, is just 0.433.

4. Concluding remarks

A software for optimal monitoring network redesign (OMNR) has been presented able to add and/or remove measurement points from an existing network. It allows one to use stochastic and deterministic approaches and to select among different objective functions (OF) covering the main desired goals of optimization. The software works in MATLAB environment and it is provided of different computational modules embedded within a graphical user interface. A case study has been presented related to the downsizing of the groundwater level monitoring network of the aquifer of Tavoliere in Apulia (South Italy). Nevertheless, many other validation tests have been performed in order to assess the software reliability and efficiency. All these tests provided excellent results. Further developments of the software have already been scheduled in order both to add new objective functions and improve the user interface.

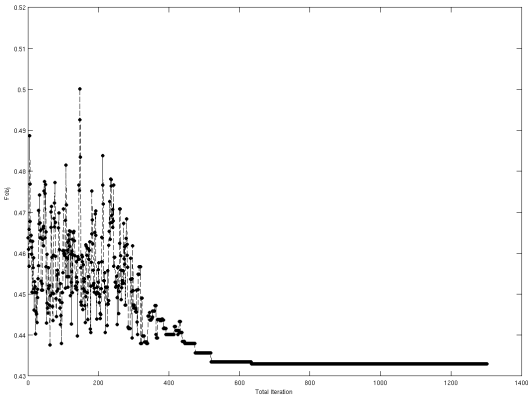


Figure 3: Behaviour of the current transitory optimal energy (*fitness value*).

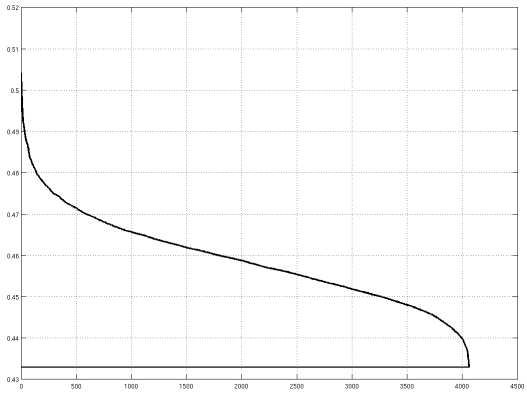


Figure 4: Values of the objective function (mean KEV) for all the 4060 possible configurations of the monitoring network.

References

Barca E., Passarella G., Uricchio V.F. (2008). Optimal extension of the rain gauge monitoring network of the Apulian Regional Consortium for Crop Protection. *Environmental Monitoring and Assessment*, 145, 1, 375-386.

Hu M.-G.,Wang J.-F. (2010). A spatial sampling optimization package using MSN theory. *Environmental Modelling & Software*, doi:10.1016/j.envsoft.2010.10.006.

Jimênez N., Toro F.M., Vélez J.I., Aguirre N., (2005). A methodology for the design of quasi-optimal monitoring networks for lakes and reservoirs. *Journal of Hydroinformatics*, 7, 105-116.

Kirkpatrick S., Gelatt C.D., Vecchi M.P. (1983). Optimization by Simulated Annealing. *Science*, 220, 671–680.

Naoum S., Tsanis I.K. (2004). Integrating multicriteria analysis and gis for assessing raingage worth within an established network. JAWRA Journal of the American Water Resources Association, 40-6, 1449–1468.

Passarella G., Vurro M., D'Agostino V., Barcelona M.J. (2003). Cokriging Optimization of Monitoring Network Configuration Based on Fuzzy and Non-Fuzzy Variogram Evaluation. *Environmental Monitoring and Assessment*, 82-1, 1-21.

Van Groenigen, J.W., Stein, A. (1998). Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality*, *27*, 1078-1086.

Van Groenigen J. W., Pieters G., Stein A. (2000). Optimizing spatial sampling for multivariate contamination in urban areas. *Environmetrics*, 11, 227-244.

Wu, Y. (2004). Optimal design of a groundwater monitoring network in Daqing, China. *Environmental Geology*, *45*, 527–535.

Comparing SaTScan and *Seg*-DBSCAN methods in spatial phenomena

Silvestro Montrone
Department of Statistics, University of Bari, s.montrone@dss.uniba.it

Paola Perchinunno
Department of Statistics, University of Bari, p.perchinunno@dss.uniba.it

Samuela L'Abbate
Department of Statistics, University of Bari, samuela.labbate@dss.uniba.it

Cosimina Ligorio
Department of Statistics, University of Bari, c.ligorio@dss.uniba.it

Abstract: The aim of this paper is to group territorial units in areas of high intensity, using SaTScan and *Seg*-DBSCAN clustering methods to aggregate adjacent spatial units that are homogeneous with respect to the phenomenon being studied. SaTScan scans the region of interest with a moving window and compares a smoothing of the intensity inside and outside it so that units belonging to contiguous windows with similar intensity are aggregated into a cluster. On the other hand, *Seg*-DBSCAN, a new version of DBSCAN, limits the arbitrariness of the choice of input parameters and identifies clusters as dense regions in space. As an application we analyze geo-referenced data concerning housing problems in Bari and we propose a comparison between the two methods presented.

Keywords: clustering, SaTScan, DBSCAN, *Seg*-DBSCAN, housing problems.

1. Introduction

Our work is prompted by the need to identify territorial areas and/or population subgroups characterized by situations of hardship or strong social exclusion through a fuzzy approach that allows the definition of a measure of the degree of belonging to the disadvantaged group. Grouping methods for territorial units are employed for areas with high (or low) intensity of the phenomenon by using clustering methods that permit the aggregation of spatial units that are both contiguous and homogeneous with respect to the phenomenon under study. This work aims to compare two different clustering methods: the first based on the technique of SaTScan and the other based on the use of *Seg*-DBSCAN, a modified version of DBSCAN.

2. SaTScan method

SaTScan scans the region of interest with a moving window and compares a smoothing of the intensity inside and outside it: units belonging to contiguous windows with similar intensity are aggregated into a cluster [2].

The identification of clusters means, therefore, to determine an area in which a set of points contributes to maximizing the incidence of the phenomenon within the area and to minimizing the incidence outside the area. In practice, the technique involves placing a monitoring window at random on the area of observation and then calculating the value of an estimator both inside and outside the area before proceeding to the testing of hypotheses.

3. *Seg*-DBSCAN method

DBSCAN (Density Based Spatial Clustering of Application with Noise) was the first density-based spatial clustering method proposed [1]. The key idea is that to define a new cluster or extend an existing cluster, a neighborhood around a point of a given radius ε must contain at least a minimum number of points *MinPts*, i.e. the density in the neighborhood is determined by the choice of a distance function for two points *p* and *q*, denoted by *dist(p,q).* The greatest advantages of DBSCAN are that it can follow the shape of the clusters and that it requires only one distance function and two input parameters [1]. Their choice is crucial because they determine whether a group is a cluster of points or a simple noise.
In order to limit the arbitrariness of the choice of a value to assign to ε, usually detected by a heuristic procedure, in this work we develop a new algorithm: *Segmented* DBSCAN (*Seg*-DBSCAN), a modified version of DBSCAN, in which the clusters are aggregated considering multiple levels of value of ε.
Therefore, to define levels of ε, a value of *MinPts* is fixed and we analyze the distribution of the maximum radius of the cores that are groups formed by *MinPts* points. Then, we build a histogram of this distribution and we choose ε where there are the histogram peaks that indicate a proximity of the cores of a cluster. As suggested in literature, we can fix the value of *MinPts* to 4, and a number of levels of ε equal to the number of the highest histogram peaks.
The final phase of the algorithm is to merge the clusters obtained. The merging of two clusters C_1 and C_2 characterized by different levels of density ε_1 and ε_2 is obtained if

$$d\left(C_1, C_2\right) < \max(\varepsilon_1; \varepsilon_2) \qquad (1).$$

With this new algorithm, parameter ε is no longer established *a priori*.

3. Distance function for application

The aim of our study is to identify the dense areas in terms of intensity compared to the considered index. For this purpose, instead of Euclidean distance a function was chosen that warps the geometric space so that points that are geographically close and have a high intensity become even closer, while points that are geographically close, but at least one of which has a low intensity, become more distant.
The function that links in these terms two points *A* and *B* of coordinates $A(x_A, y_A, w_A)$ and $A(x_B, y_B, w_B)$ respectively, with $0 < \{w_A, w_B\} < 1$, is a weighted distance that is obtained by dividing the Euclidean distance by a mean of order integer $t > 0$:

$$d_{pesata}(A,B) = \frac{\sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}}{\sqrt[t]{\left(\frac{w_A^{-t} + w_B^{-t}}{2}\right)^{-1}}} \qquad (2).$$

Observe that in this distance the triangle inequality does not hold, so it is a semimetric, but this restriction does not affect the definitions of density-reachability and density-connectivity necessary for DBSCAN algorithm [1].
With this function the distance increases in matching pairs of points with low intensity value, so that they are penalized in the formation of clusters. Empirically it was verified that the most appropriate value of t is 5.

3. Application

This work aims to identify the land areas characterized by situations of housing problems by defining typical indicators able to estimate the difficulty in small areas. The case study uses data from the last Population and Housing Census carried out by ISTAT in 2001. The indices were calculated for each section of the census of the city of Bari [3]:

- incidence of the number of dwellings occupied by rent-payers with respect to the total number of dwellings occupied by residents;
- index of overcrowding: the ratio between the total number of residents and size of dwellings occupied by residents;
- availability of functional services: landline telephone, the presence of heating systems and the availability of a designated residential parking space.

These indices may be synthesized by a fuzzy index obtained by "Total Fuzzy and Relative" (TFR) method [3]; we denominate this new index “disadvantaged housing index”. It is a measure of an individual’s degree of membership to a disadvantaged group and its range is between zero (if the individual does not definitely belong to this group) and one (if the individual definitely belongs to this group).
Using the SatScan method, we identify different clusters each composed by a different number of sections of the city of Bari.
The city of Bari presents various critical areas: the old town of San Nicola, the areas surrounding the city center, Madonnella, Libertà and Carrassi (the former characterized by the presence of public housing complexes such as the Duca degli Abruzzi). Less critical, though more widespread, is the situation in some suburban areas such as Carbonara and Ceglie.
The same data on housing problems were analyzed with the *Seg*-DBSCAN method by associating geographic coordinates to the disadvantaged index to obtain eight clusters. The critical areas thus obtained do not exactly coincide with those identified by the SatScan method: both methods identified the old town of San Nicola and the areas surrounding the city center - Madonnella, Libertà and Carrassi – as well as Carbonara and Ceglie; but the *Seg*-DBSCAN method also identified the districts of San Cataldo and San Paolo
We observe that SatScan identifies areas formed by contiguous spatial units in which a smoothing of the disadvantaged housing index is performed. This method is effective in

identifying areas of high or low intensity and therefore may be a useful indication of areas "at risk" to be monitored.
Like the SatScan method, *Seg*-DBSCAN identifies areas in which the spatial units meet a criterion of adjacency, but *Seg*-DBSCAN differs in excluding those areas where the phenomenon is absent. *Seg*-DBSCAN can exactly identify sections of the city with housing problems. In the case of the San Nicola district, the old town of Bari, the SaTScan method identifies the whole district (Figure 1a) while the *Seg*-DBSCAN method identifies the same area of hardship but also analyzes the area in more detail (Figure 1b). The method identifies the particular points with a greater presence of the phenomenon and excludes the points where the phenomenon is not present because of the restoration of historic buildings.

Figure 1a: SaTScan method **Figure 1b**: *Seg*-DBSCAN method

4. Conclusions and future advancements

The proposed methodologies identify areas where there is a high disadvantaged index. As we have noted above, a comparison of the two methods shows that the *Seg*-DBSCAN method is more accurate in identifying the spatial units in which there are housing problems. The future advancement of our work will be to seek a cluster validity index for spatial data, which takes into account the noise points, that is valid from a statistical point of view and that allows the accurate measurement of the *Seg*-DBSCAN method.

References

[1] Ester M., Kriegel H.-P., Sander J., Xu X. (1996) *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland.

[2] Kulldorff M.(1997) *A spatial scan statistic* Communications in Statistics Theory and Methods 26(6) 1481-1496.

[3] Montrone S., Perchinunno P., Di Giuro A., Rotondo F., Torre C.M. (2009) *Identification of "Hot Spots" of Social and Housing Difficulty in Urban Areas: Scan Statistics for Housing Market and Urban Planning Policies* in: *Geocomputation and Urban Planning*, Murgante B., Borruso G., Lapucci A. (Eds.), Springer, 57-78.

Fire, earthquake, landslide, volcano, flood: a first approach to a natural hazard map of Italy

Rina Camporese, Niccolò Iandelli
Iuav University of Venice, rina.camporese@gmail.com, niccogeo@gmail.com

Abstract: Several natural hazards have been synthesized in one map to obtain an overall assessment of these phenomena. Space and time coherent data have been searched for: minimum disaggregation available was the province and matching year 2007. Two indices of susceptibility have been calculated for fires and landslides; for seismic hazard the median value of provincial values of maximum ground acceleration has been used. For each index, provinces have been classified in four quantile levels. A weighted average of the three classified levels has been calculated with weights proportional to annual expenditure for every type of event. Results are meaningful at ordinal scale and cannot be interpreted as a measure of risk. Floods and volcanoes have been mapped, too, thus obtaining a global overview of the main natural hazards in Italy.

Keywords: natural hazards, integrated data, composite indicator, overview map

1. Introduction

In a future scenario, territorial information systems could provide an interactive map which crosses vulnerability caused by several natural hazards and the value represented by the population and human artifacts, outlining the concept of "risk". It is possible to figure out a complex integration of hazard maps, vulnerability maps and value maps to represent the spatial distribution of risk, enabling, for example, an individual citizen to evaluate the risk of being in different places. The foundations of this work lay in the answer to a few questions: to what extent can data on various natural hazards be integrated? Do current knowledge and tools enable to build an integrated view of such risks? How realistic and significant can a synthetic measure be?
A first step towards a risk map is the integration of the various hazards. This work aims at providing an integrated framework of natural hazards, through authoritative sources, available at Italian national level in a consistent way, both in space and in time alike. It has been inspired by some previous research aiming at synthesizing different environmental hazards into one global measure (Arnold et al. 2007, European Commission 2007, ISPRA 2008).

2. Materials and Methods

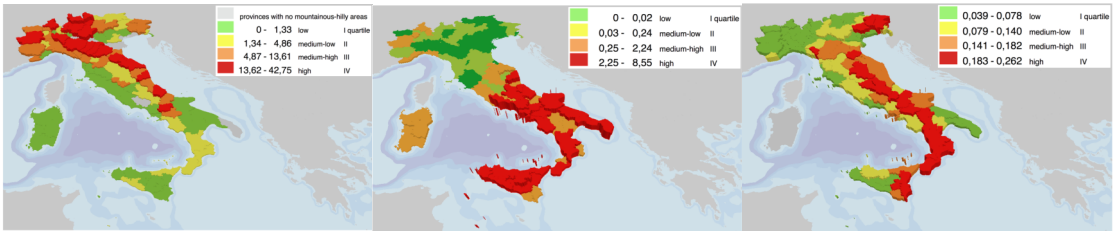
Data are related to very different phenomena and parameters. The minimum territorial unit has been forcibly the province, because, for some phenomena, that was the highest level of detail available for the entire territory. Given the difficulty of correlating data so different in nature and spatial trend, a synthetic index of dangerousness for each type of event has been created. Data are described in detail here below.
▪ *Landslides*: surface of landslide areas in 2007 are published by the National Institute

for Environmental Protection and Research[1]. The ratio between the landslide area and the surface of the mountain-hill area[2] for each Province has been calculated: it represents the quote of landslides in the areas potentially affected by landslides.

▪ *Forest fires*: data on forest areas and burnt forests in 2007, published by the Fire Service of the State Forestry Body, enabled the calculation of a similar index: the percentage ratio between the wooded area affected by fires and the total wooded area in each province.

▪ *Earthquakes*: In 2006 the National Institute of Geophysics and Volcanology published data on seismic hazard in terms of maximum ground acceleration with 10% exceeding probability in 50 years, referred to bedrock, calculated on a grid of points, with a step of 0.02 degrees They measure the acceleration at which the ground is expected to be exposed and are not a measure of seismic risk, which should include also the losses caused by earthquakes, in terms of direct casualties and damages. The 55,689 points on the mainland have been attributed to the corresponding province, except for Sardinia where data were not available. The median of the values of seismic hazard within a province has been used as a synthetic provincial measure. The choice has been made after a thorough exploratory data analysis with the aim of identifying a single summary measure enabling to sort the provinces on the basis of seismic hazard, similarly to what has been done for landslides and fires. However, the size and shape of the provincial administrative areas are such that many provinces have highly variable values (e.g. the province of Reggio Calabria has a bimodal distribution). In such situations, the characterization of the entire province by the median value can be misleading, since the intra-provincial variability is high and it is linked to geo-structural factors, such as capable faults and geodynamic activity.

The indexes for landslides and forest fires are both composition ratios, perfectly comparable in general terms, except for the reference period: for landslides it has its upper limit in 2007, but it also includes landslides originated in previous years and still active, while for forest fires only areas burned in 2007 have been considered. The indicator of seismic hazard has a completely different nature: it represents the median value of ground acceleration within a province.



Landslides Forest fires Earthquakes

Figure 1: Hazard indexes by province - Italy, 2007

To bring the indicators back to a common scale, for each of the three hazard indices, provinces have been classified according to their position in the national ranking with regard to distribution quartiles: 25% of the provinces with low values, 25% medium-low, 25% medium-high, 25% high values (fig. 1). This can be viewed as an extension of the normalization method for indicators above or below the mean.

Finally, a synthetic measure of the joint dangerousness of the three events has been calculated in order to obtain a single integrated map of the various hazards: it has the

[1] IFFI project - Inventory of landslides in Italy.

[2] The mountainous-hilly areas have been calculated by using National Institute of Statistics data on Italian municipalities at 2009.

advantage of offering a synthetic view of the hazards for all the natural disasters considered at provincial level.
To take into consideration the different impact of events in terms of damage, estimated annual costs incurred in Italy because of these events have been used as weighs. In other words, to each of the three natural disaster index a weight has been assigned, proportionally to the severity of the outcomes, assessed in billions of euros of damage per year. For forest fires, an estimate of € 0.6 billion[3] spent in 2007 has been used, for earthquakes a cost of € 3.4 billion[4] and for landslides the value of € 1.5 billion[5] per year have been estimated. Weights, interpretable as relative severity coefficients, have been calculated as the ratio between the annual cost per event (landslides or earthquakes) and annual expenditure for fire events, i.e. the less expensive event. These are the values: Fire-weight = 1, Landslide-weight = 2.5, Earthquake-weight = 5.7. In order to differentiate the values and widen the range of provincial indicators, the sum of weighted values of the three indices has been chosen. The formula is:

$$[1] \qquad H = Fw * F_{ire}\text{-}i + Lw * L_{andslide}\text{-}i * + Ew * E_{arthquake}\text{-}i$$

The implication connected with linear aggregation, i.e. that there are no synergies or conflicts among the different aspects considered, seems acceptable. The composition method is elementary, one could say rough; this is because, at this preliminary stage of the research, the desired output didn't want to be a ranking of provinces, but rather a classification of provinces into four ordinal classes to be mapped.
To complete the picture, floods recorded in 2007 and areas potentially affected by volcanic eruptions have also been represented on the map. Volcanoes are concentrated in few areas of the country and it would have been pointless to calculate the danger in all the provinces. For floods, unfortunately, it was not possible to calculate an indicator similar to the others, since no areal pieces of information on affected and potentially affected zones were available for the entire country.

3. Results

This first synthetic map of natural hazards in Italy highlights the danger deriving from the relatively young geological age of our peninsula. High values of the synthetic index are observed along the Apennine ridge and in alpine areas still tectonically active, i.e. Friuli Venezia Giulia in the north-east. The map is strictly connected to the morphology and the active geodynamic processes and it identifies the southern provinces as characterized by high hazard levels, since they are subject to dangerous natural phenomena with high destructive potential (earthquakes, landslides, volcanoes). Such phenomena have different degrees of freedom, and often do not occur in a cyclical way, i.e. times of recurrence cannot be calculated or predicted. The synthetic index draws a geography of hazard that may be useful for planning actions or for directing resources aimed at mitigating the effects of such phenomena. Unfortunately, the level of data disaggregation (province) is not adequate to identify any smaller "black spot".

[3] A study conducted in Spain by WWF estimated a cost of 5,500€ per hectare of forest burnt. As in 2007 in Italy 116,602 hectares were burnt, the total annual expenditure incurred due to fires is estimated to be €0.6 billion.
[4] According to the Italian Civil Protection "earthquakes which struck the peninsula [Italy] have caused substantial economic losses, assessed for the last forty years in approximately 135 billion euros".
[5] According to data coming from the Census of landslides from 1918 to 1994 (32,000 landslides surveyed) conducted by the National Research Council (CNR) - Project GNDCI AVI - the damage caused each year amounted on average to 1 or 2 billion euro.

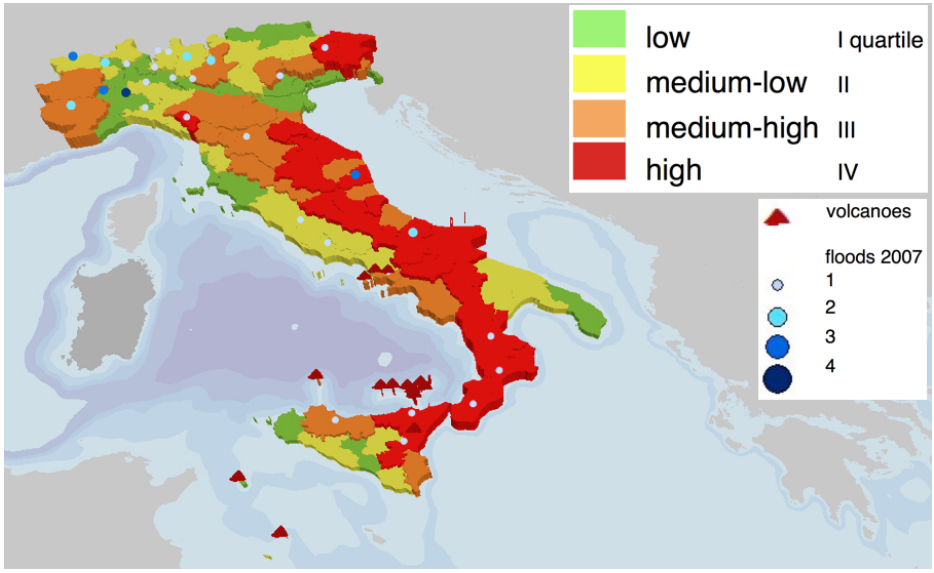


Figure 2: Synthetic natural hazard index by province - Italy, 2007

4. Concluding remarks

The calculation method and the system of weights are subjective and only enable to arrange provinces in a rank made of four values. The result does not allow a fine evaluation of the global hazard in quantitative terms, since it is significant only in ordinal scale, i.e. the numerical differences between the indices of two provinces are not significant, neither can this be interpreted as a measure of risk. Furthermore, the expected period of recurrence of the phenomena is not taken into account. Alternative normalization, weighting and composition methods could be used to evaluate the performance of different composition procedures (OECD, EC, JRC, 2008); Multi Criteria Evaluation (MCE-GIS) could be performed, too (Chen et al., 2001). Anyway, beyond technical statistical issues, the major point here seems to be the scarce availability of spatially detailed, comparable and timely information. Nevertheless, despite these limitations, the final map offers a synthetic view of the natural hazards at provincial level and has the advantage of a comprehensive look at all the natural disasters taken into consideration - fires, earthquakes, landslides, volcanoes, floods - based on standardized methods for the entire country.

References

Arnold M., Dilley M., Deichmann U., Chen R.S., Lerner-Lam A.L. (2005) *Natural Disaster Hotspots: A Global Risk Analysis*, The World Bank.

Chen K., Blonga R., Jacobson C. (2001) MCE-RISK: integrating multicriteria evaluation and GIS for risk decision-making in natural hazards, *Environmental Modeling & Software*, Vol. 16, Issue 4, p. 387-397.

European Commission (2007) *Armonia: Assessing and Mapping Multiple Risks for Spatial Planning - Approaches, methodologies and tools in Europe*, Lancaster University, Department of Geography.

ISPRA Institute for Environmental Protection and Research (2008) *Environmental Data Yearbook – Natural and Anthropogenic Hazard*.

OECD, European Commission, Joint Research Centre (2008) *Handbook on constructing composite indicators: methodology and user guide*, OECD Publishing.

Spatio-temporal analysis of wildfire patterns in Galicia (NW Spain) [1]

Isabel Fuentes-Santos, Wenceslao Gonzalez-Manteiga [1]
Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, isabel.fuentes@usc.es

Manuel F. Marey-Pérez [2]
Departamento de Ingeniería Agroforestal. Universidad de Santiago de Compostela.

Abstract: In this work some of the analysis and inference techniques developed recently for spatial point patterns are applied in order to analyze spatial patterns of wildfire ignitions recorded in Galicia (NW Spain) in the period 1999-2008.

Keywords: Spatial point process, intensity, stationarity, K-function, wildfire ignition point.

1. Introduction

Spatial point patterns arise in a wide variety of scientific contexts, including seismology, forestry, geography and epidemiology, (Diggle, 2003).
Wildfire is the most ubiquitous natural disturbance in the word and represents a problem of considerable social and environmental importance. In this work we analyze the spatio-temporal pattern of wildfire ignitions in Galicia (NW Spain), where arson fires are the main cause of forest destruction, in order to model and predict fire occurrence. Such information is of great value in elaborating fire prevention and fire fighting plans.

2. Materials and Methods

Data set:
In this study, the spatio-temporal pattern of wildfires recorded in Galicia during the period 1999-2008 is analyzed. Galicia is located in the North-West of the Iberian peninsula and has a surface area of 29,574 km^2 (11,419 sq mi), which 69% is covered by forests. The total number of fires recorded in the study area from 1999 to 2008 is 85,134. In addition to the spatial location and the date of occurrence of the ignition points, we consider two marks: cause (arson (82.5%), natural, negligence, reproduction and unknown cause) and the size of the burned area.

Statistical methods.
A spatial point process X is a stochastic model governing the locations of events $\{x_i; i=1,...,n\}$ in a bounded region $A \subset R^2$ (Diggle 2003). If the point process contains

[1] (MTM2008-03010)

associated measures or marks, it is referred as a marked point process. A point process is characterized by the probability function $P\left(N\left(A\right)=N\right)$, where $N\left(A\right)=\#\left(x_i \in A\right)$, which is the probability of finding N events in the region A, and by its first and second order characteristics. The (first order) intensity, ***λ(x)***, is the point process analogue to the mean function for a real-valued stochastic process. Second order characteristics describe the spatial structure of point processes and are based on the analysis of pairs of points. Although several second order characteristics have been developed to describe point patterns (Diggle, 2003), this work focuses on the analysis of the reduced second order moment measure or Ripley´s K-function. (Ripley, 1977), which expresses the expected number of events within a ball b(x,r) centred in an arbitrary event x. The point process is stationary and isotropic if its statistical properties do not change under translation and rotation, respectively. Under these conditions the intensity function is a constant λ, equal to the expected number of events per unit area. In the non-stationary case, the intensity depends on the individual locations.

The first step in the analysis of an observed spatial point pattern is to test the complete spatial randomness (CSR) hypothesis, under this assumption the data are a realization of a homogeneous Poisson process, which is characterized by two properties: (i) the expected number of events (fires) in a flat area (study area) $A \subset R^2$ of surface area |A| has Poisson distribution with mean λ|A|, and (ii) for n events $\{x_i, i=1,\ldots,n\}$ in A, these are a random sample of the uniform distribution in A. The constant λ in (i) is the intensity of the process. According to (ii), there are no interactions between events (Poisson). This property acts as a dividing hypothesis between regular and aggregated patterns.

In this work, the stationarity assumption was tested by the measure of inhomogeneity proposed by Comas *et at*. (2009):

$$\hat{S} = \int_A \left\| \hat{\lambda} - \hat{\lambda}\left(x\right) \right\| dx \tag{1}$$

where $\hat{\lambda} = N/A$ and $\hat{\lambda}\left(x\right)$ the non-parametric kernel estimator of the intensity (Diggle, 1985).

The second order structure of the observed patterns was characterized by the estimate of the inhomgeneous K-function proposed by Baddeley et al. 2000:

$$\hat{K}_{in\hom}\left(r\right) = \frac{1}{|A|} \sum_{x_i \in X \cap A} \sum_{x_j \in (X \cap A)\setminus\{x_i\}} \frac{I\left(\left\|x_i - x_j\right\| \le r\right)}{\hat{\lambda}\left(x_i\right)\hat{\lambda}\left(x_j\right) w_{ij}} \tag{2}$$

where w_{ij} is Ripley´s edge correction factor. Specifically, a Monte-Carlo test based on the inhomogeneous L-function $L_{in\hom}\left(r\right) = \sqrt{K_{in\hom}\left(r\right)/\pi}$ was applied to test the Poisson hypothesis, since it is easier to visualize dependence between points, as for a Poisson process $L_{inhom}(r)=r$.

The wildfires database is a spatial point process marked by cause and size of the burned area. Spatial interaction between events of two types occurs when different types of events are either closer or further apart than expected under independence. This hypothesis is tested applying a Monte Carlo test based on the inhomogeneous L-cross function. For ease of comparison the L-index (Genton et al. 2006), that enables presentation of the test for several pairs of patterns in a single plot, and its simulation envelopes were computed to analyze the spatial dependence between ignition points in pairs of sequential weeks.

3. Results

As described in section 2, the dataset contains the ignition points of wildfires reported in Galicia during the period 1999-2008, marked by cause and size of burned area. In this section we present some results of the analysis of wildfires recorded in the whole period and, for ease of interpretation, wildfires recorded in 2006.

In order to characterize the degree of inhomogeneity of the different patterns, the maximum value ($\hat{S}_{\max}$) of $\left\{\hat{S}_b^*, b=1,...,B\right\}$, for B=20 realizations of a homogeneous Poisson process involving the same number of fires as the observed pattern, was compared with the empirical $\hat{S}$ for the original pattern. When $\hat{S} > \hat{S}_{\max}$, we reject the stationarity assumption. This test shows that all the patterns analyzed should be considered non-stationary, see results for fires classified by cause in table 1. The kernel intensity estimates for these patterns (figure 1) confirms the importance of arson fires in Galicia and shows that the South and South-West of the region are the most conflictive areas, expect for natural fires, which present higher intensity in the East.

	Fires	$\hat{S}_{obs}$	$\hat{S}_{\max}$
Arson	70223	41939.7	1667.0
Natural	887	526.6	101.0
Negligence	4224	1707.6	255.2
Reproduction	2574	2224.4	242.6
Unknown	7226	4844.7	423.7
Total	85134	48161.2	1707.8

Table 1: Stationariy test for wildfires recorded in 1999-2008.

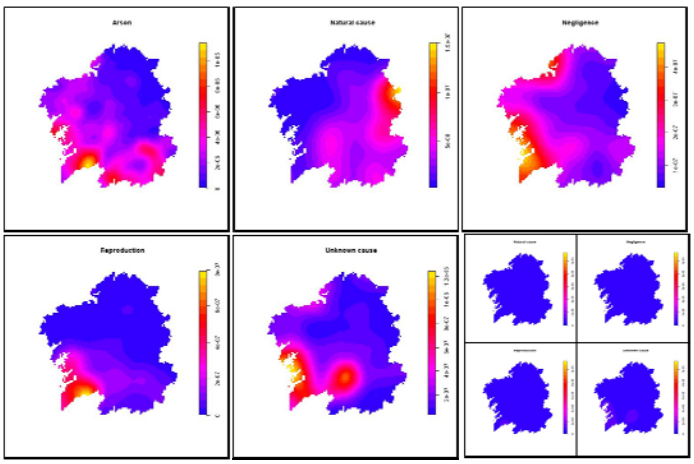


Figure 1: Kernel intensity estimate for wildfires by cause in Galicia 1999-2008. Bottom right: comparison between arson fires and rest of causes.

Independence L-tests applied to some spatial patterns of wildfires recorded in 2006 are shown in figure 2. Comparison of inhomogeneous L-tests (top left and middle) shows more evidence of aggregation for arson fires than for natural fires. The inhomogeneous

L-cross (top right) shows positive interaction between small and large fires up to 6 km. Finally, the L-index test shows positive spatial interaction between wildfires in consecutive weeks assuming both homogeneous and inhomogeneous patterns, although the evidence is higher for the homogeneous test.

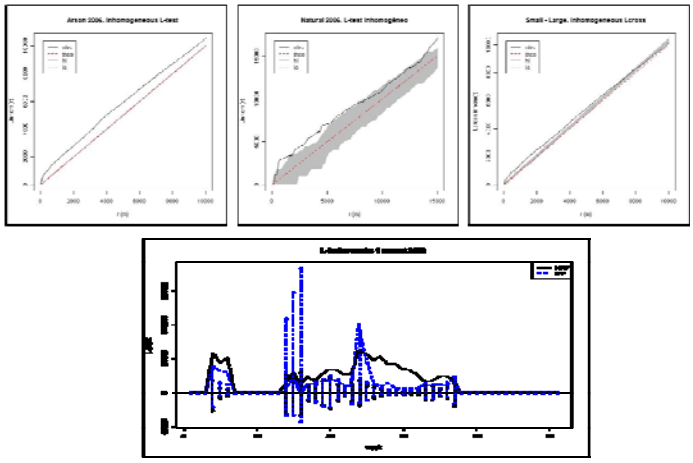


Figure 2: Second order analysis 2006. Top left: Inhomogeneous L-test arson fires; middle: inhomogeneous L-test natural fires; right: L-cross small-large fires. Bottom: L-index for consecutive weeks.

4. Concluding remarks

In this work we have seen the utility of spatial point processes in the analysis of wildfires.

Taking into account the results obtained, we propose to include meteorological and socioeconomic variables in order to fit an accurate spatial model. Finally, we propose to consider the spatio-temporal point pattern defined by spatial location and starting date of wildfires, test for separability and fit a spatio-temporal model.

References

Baddeley A.J., Moller J, Waagepetersen R (2000) Non and semi- parametric estimation of interactions in inhomogeneous point patterns. *Statistica Neerlandica*, 54, 329-50.

Comas, C., Palahi, M., Pukkala, T., Mateu, J. (2009). Characterising forest spatial structure through inhomogeneous second order characteristics. *Stochastic Environmental Research and Risk Assessment*, 23 (3), 387-397

Diggle P.J. (2003). *Statistical Analysis of Spatial Point Patterns.* Oxford University Press.

Genton M.G., Butry D.T., Gumpertz M.L., Prestemon, JP (2006) Spatio-temporal analysis of wildfire ignitions in the St Johns River Water Management District, Florida. *International Journal of Wildland Fire* 15, 87-97.

Ripley BD (1977) Modelling spatial patterns (with discussion). Journal of the Royal Statistical Society B 39, 172

Imputation strategy in spatial data

Laura Martino and Alessandra Palmieri
European Commission – DG ESTAT
e-mail: alessandra.palmieri@ec.europa.eu

Abstract

In area frame surveys a mixed approach consisting in the observation of the surroundings of a not reachable point combined with orthophoto interpretation needs to be used in the locations that are particularly remote and difficult to reach. In this situation a simplified nomenclature has to be applied for some land cover categories due to the difficulties in properly distinguishing among specific classes. In the estimation phase the resulting observations can be considered affected by a sort of partial non response phenomenon. Classification of land cover indeed is available only at aggregated level. A donor based methodology is proposed to impute this missing detailed information. Assuming that neighbouring points are affected by spatial autocorrelation, potential sets of donors are identified among points within different distance thresholds. Capability of the method of correctly imputing the missing information is discussed and its robustness assessed in terms of two different cost-functions both based on the maximum distance observed among potential donors and recipient points.

Keywords: area frame survey, missing data, hot deck imputation, spatial data

1. Introduction

Area frame surveys usually foresee people going to the field and colleting in-situ information that are visible on the ground. This could be the case of crops, environmental parameters, forestry features and so on.
Since the accessibility to the point can be difficult for many reasons (fences, military areas, wild animals, etc.) it could be the case that for some units it is impossible to assess the land coverage in-situ. In these situations the recourse to a mixed approach - consisting in the observation of the surroundings of the point combined with orthophoto interpretation - is frequently adopted. As a consequence a simplified nomenclature needs to be applied for some land cover categories due to the difficulties in properly distinguishing among specific classes (i.e. durum wheat from oats and barley). In the estimation phase the resulting observations can be considered affected by a sort of partial non response phenomenon (some detailed information on land cover is missing). Classification of land cover indeed is available only at aggregated level.
Various methodologies are available to cope with partial missing data information (Little & Rubin, 1987). When data are affected by spatial correlation, the location of the sampling units can play an important role in the prediction of the missing information.

2. Imputation strategy for an area frame survey

One of the methodologies most commonly used to cope with missing data issues is the hot deck imputation (King C. S. & Bogle R. D., 2003, Gabriella Schoier, 1999). It consists of filling in missing values on incomplete records using values from similar, but complete records of the same dataset or external dataset. The identification of the best donor for each incomplete record can be based on different criteria like distance function matching or nearest neighbour. When spatial data are considered and sampling units are portion of land, physical distance among points usually represents a good indicator of similarity. Nonetheless to guarantee robustness of the imputation procedure, techniques taking into consideration the distribution of the set of donors need to be considered.

A methodology is proposed here taking into consideration both the need to look at the distribution of the land cover classes among the donor sets and the minimization of an overall indicator of distance between donor and recipient point.

The main steps of the methodology for each point affected by partial missing information are the following:

- Five distance thresholds are defined (10, 15, 20, 25 and 30 km);
- five nested sets of donors are set up composed of all the points belonging to the same stratum and lying progressively further off the recipient point (in a circle of ray equal to the threshold distance);
- a sort of 'standardized modal value' of the distribution of each set of donors is computed standardizing the relative frequency of each land cover class by the general share of the corresponding land cover in the country;
- the best donor set/value among those previously set up is selected maximizing a gain function;
- the modal value of the selected donor set is attributed to the recipient point.

2.1 The standardized modal value

The standardized frequency of each land cover in a donor set is computed dividing the relative frequency of each land cover in each donor set by the corresponding share in the population

$$\hat{r}_{L,s} = r_{L,s} / r_L$$

Where

$\hat{r}_{L,s}$ standardized relative frequency of the Land Cover L-th in the donor set s-th

$r_{L,s}$ relative frequency of the Land Cover L-th in the donor set s-th

r_L relative frequency of the Land Cover L-th in the population

The standardized modal value is the land cover class having the highest standardized relative frequency.

This device was introduced to avoid that the donor value was biased in favour of land cover classes that have the largest share in the general population.

2.2 The gain functions

Two gain functions are proposed both linked to the maximum distance of the points belonging to each set of donors (ray of the circle) and the modal frequency of the land cover observed on the points belonging to each donor set. The aim of these functions is to favour the choice of the donor value most frequently found in the closest surroundings (measured as absolute distance or area) of the recipient point.

The first function is expressed as the ratio of the modal frequency and the area of the circle centred on the recipient point. It express how typical is the modal value in the area of circular shape surroundings the recipient point.

$$G_s = \hat{r}_{M_s} / \left(\left(Maxd_{M_s} \right)^2 * \pi \right)$$

The second cost function is based on the linear distance. It provides a measure on how further it is needed to go to find donor-value representative of the area.

$$H_s = \hat{r}_{M_s} / Maxd_{M_s}$$

Where:

k= $1,\ldots, s, \ldots, n_s$ set of donors satisfying the conditions

1) $\{ i \in s : i \in s+1, ..., n_s \}$

2) $\{ \forall i \in s, \; j \in s+1 : d_i < d_j \}$

M_s modal land cover class of the distribution of the s-*th* set of donors;

$\hat{r}_{M_s}$ standardized frequency of the standardized modal land cover class of the distribution of the s-*th* set of donors;

d_{M_s} distance of the donors having the modal land cover class from the recipient;

$Maxd_{M_s}$ maximum distance from the recipient point of the donors having the modal value.

3. A case-study: the European Land Use and Cover Area frame Survey (LUCAS)

The capability of the method of correctly input missing data is tested on the European 2009 LUCAS survey. The LUCAS (Land Use/Cover Statistical Area Frame Survey) survey is a field survey based on an area-frame sampling scheme (Martino & Fritz, 2008). Data on land cover and land use are collected and landscape photographs are taken. Eurostat carried out the largest ever LUCAS campaign in 2009. It collected data on the ground on land cover, land use and landscape diversity on approximately 234,000 points. Those points were selected from a standard 2 km grid with in total 1 million points all over the EU. The land cover and the visible land use data were classified according to the harmonized LUCAS land cover and land use nomenclatures.

The complete records of the LUCAS 2009 (http://epp.eurostat.ec.europa.eu/portal/page/portal/lucas/data/database) have been used to test the capability of the method of properly imputing the missing data through a simulation exercise. Starting from the complete set of points with arable and permanent

crops in Europe (46,296 out of 234,907) the true land cover value of a single point has been deleted, one by one, and all the other points in the same country/stratum (arable land or permanent crop) have been treated as potential donors. Then distances between the recipient point and the others have been computed and five nested sets of donors defined according to the thresholds of 10, 15, 20, 25 and 30 km respectively. The new land cover category is imputed on the basis of the proposed methodology.
Some quality indicators have been computed to evaluate:

1. The capability of the methodology of imputing the correct value (unbiasness) by land cover (with 28 and 7 classes) and by country. This indicator is expressed as the percentage rate of agreement between the imputed and the true value;
2. The robustness of the obtained results with respects of the different distance thresholds and gain functions. This is expressed as the number of times the same land cover class is imputed out of the five potential sets of donors and with respect of the two different gain functions.

All the countries surveyed in 2009 have been included in the simulation. Their diversity in terms of land cover landscape (expressed as Shannon Evenness Index) has been accessed and analyzed in combination with the quality of the results to better understand whether it could be an important factor to improve the quality of the simulation.

The overall rate of agreement is not significantly different using the two gain functions ranging between 41% and 72% (depending on how detailed is the nomenclature adopted. See Table 1) for gain function 1 and between 42% and 73% for gain function 2. A large variability is observed at country level although.

Table 1: Overall rate of accordance with true land cover

	Gain function 1		Gain function 2	
	n.	%	n.	%
	Nomenclature 2 digit			
Disagreement	31960	59	31600	58
Agreement	22286	41	22646	42
	Nomenclature 1 digit			
Disagreement	15138	28	14688	27
Agreement	39108	72	39558	73

The set of donor with the smallest size (10 km distance) seems to be the preferred one when using the gain function 1, while the largest set (30 km distance) is the one providing donation most frequently when it goes to the second gain function.

References

Little R.J.A. & Rubin D.B. (1987) *Statistical analysis with missing data.* Wiley, New York.

Gabriella Schoier (1999) *On partial non response situations: the hot deck imputationMethod.* ISI99, Helsinki 10-18 August 1999, Finland

King C. S. & Bogle R. D. (2003) *Using Hot Deck Donor Imputation Methodology in the Service Annual Survey.* ASA 2003 Joint Statistical Meetings - Alexandria, VA, US

Martino L. & Fritz M. (2008) New insight into land cover and land use in Europe, *Statistics in Focus,* 33, Eurostat, Luxembourg

Multivariate geostatistical model to map soil properties at a region scale from airborne hyperspectral imagery and scattered soil field surveys: dealing with large dimensions

Monestiez P.[1], Walker E., Gomez C., Lagacherie P.
Biostatistics and Spatial Processes, INRA, France; e-mail: monestiez@avignon.inra.fr

Abstract: Recent developments in soil sensing technologies, initially oriented towards soil mapping at the field scale for precision agriculture, show high potential for digital soil mapping (DSM) of large areas. We present here a spatial statistical model that combines hyperspectral remote sensing, field measurements and, potentially soil types from existing pedological maps, to predict soil properties as clay or calcium carbonate contents at increasing resolutions from 5m to 100m over large regions. Methodological difficulties arise from dimensional aspects. From a spatial point of view, the geostatistical model have to be inferred from rare field soil samples and remote sensing data that are patchy - only informative on bare soils - and very numerous - several thousand records at fine resolution. From a multivariate point of view, soil properties have to be predicted using PLS from high dimensional – 256 bands – hyperspectral data. To illustrate the proposed approach, a 25-square-km area located in the vineyard plain of Languedoc was surveyed with both airborne hyperspectral remote sensing data at a 5-m resolution and a survey of 200 points with soil measurements. Various maps of clay and calcium-carbonate content were produced by block co-kriging and represent different compromises between prediction accuracy and spatial resolution.

Optimal location and size for a biomass plant: application of a GIS methodology to the "Capitanata" district

Cammerino A.R.B., lo Storto M.C., Monteleone M.
Department of Agro-environmental Science, Chemistry and Plant Protection
University of Foggia, Via Napoli 25 Foggia – Italy
m.monteleone@unifg.it

Abstract: The aim of this work is to outline a methodological procedure to assess the most advantageous logistic use of agricultural residues (such as straws) in order to supply a biomass energy plant using the economic criterion of the maximization of the NPV (Net Present Value). A GIS (Geographic Information System) neighborhood statistics procedure was applied in order to locate the biomass plant.
Results showed that the optimal radius of the supply basin was not related to (but independent of) both total amount and spatial distribution of biomass resources within the basin; differently, biomass availability strongly affected the size of the plant and the corresponding NPV. Therefore, the optimal plant location was at the center of the geographical area constantly characterized by the highest biomass density at different orders of scale.
***Keywords**: straws, biomass plant, optimization, GIS, neighborhood statistics.*

1. Introduction

"Capitanata" is a geographical area of the Apulia region with a very large availability of straws, agricultural residues obtained from winter cereal crops. The total area considered in this study takes into account a part of "Capitanata" called "Tavoliere" and comprises neighbouring districts that belong to three different southern regions of Italy: Puglia, Basilicata and Campania, respectively (*Fig.1, A*); within a total area extended 665,000 hectares, 400,000 hectares are cultivated with winter cereals (average surface fraction F=0.6) from which 480,000 tons of straws are annually potentially produced.
As the size of the power plant (P) increases also the total amount of electricity produced and sold increases; however, the greater amount of feedstock needed to satisfy plant demand requires greater transportation distances, thus increasing the total hauling costs (*Leboreiro and Hilaly, 2011*). As a result of these competing factors, an optimal radius of the supply basin and an optimal plant size which maximize the profitability of the investment should be detected. The optimum plant size is also significantly impacted by the economies of scale; for the sake of simplicity, we have not considered this aspect in the present short paper. The economic criterion applied to reach these "optimal" solutions is to maximize the NPV (Net Present Value) of the overall project investment. We were specifically interested in defining three main features: 1) the "optimal" radius of the biomass supply basin (R*); 2) the "optimal" geographic location of the plant within the same basin; 3) the "optimum" size (or capacity) of the plant (P*).
Land planning in the bioenergy sector requires the processing of geo-referenced data, with particular emphasis on the spatial distribution of the available biomass resources (*Rozakis et al., 2001a*). On this respect, GIS software applications are an essential tool to work out spatial analysis from digital maps of land use and of the road network.

2. Materials and Methods

2.1 Preliminary spatial procedures. The spatial analysis was performed using ESRI ArcGIS software package 9.1. The database employed is the CASI-INEA land use map (2001); the land class "non-irrigated arable land" is strictly related to the presence of winter cereal crops from which straws derive.
Firstly, a regular grid (each cell being 2,000 x 2,000 meters corresponding to *S*=400 hectares) was overlapped to the vector land cover map so that it was possible to estimate the wheat surface fraction (F_i) in each reference unit (*cell*). The available straw per unit of cultivated area (*Y*) was estimated equal to 1.2 t ha^{-1} y^{-1} of dry biomass. The "biomass map" was then obtained multiplying, in each cell, *S* by *Y* and by F_i (*Fig.1, B*). Secondly, taking into account the provincial and national road networks, downloaded from the National Cartographic Service, it was possible to compute the distance between whichever hypothetical plant location and the centroid of each cell belonging to the whole area under study (*Alfonso et al., 2009*) so that the transportation cost of the total biomass could be determined.

2.2 Calculation of NPV. To calculate the NPV, the revenues (*Rev*) and total costs (*Cst*) regarding the annual plant operation were determined. The resulting difference (*Rev-Cst*) is the "net benefit", a constant annual cash flow that is financially brought back to the starting year of investment, applying a discount factor. Subtracting from this discounted capital the initial investment, the NPV is obtained; it represents the net profitability resulting from the overall activity undertaken.

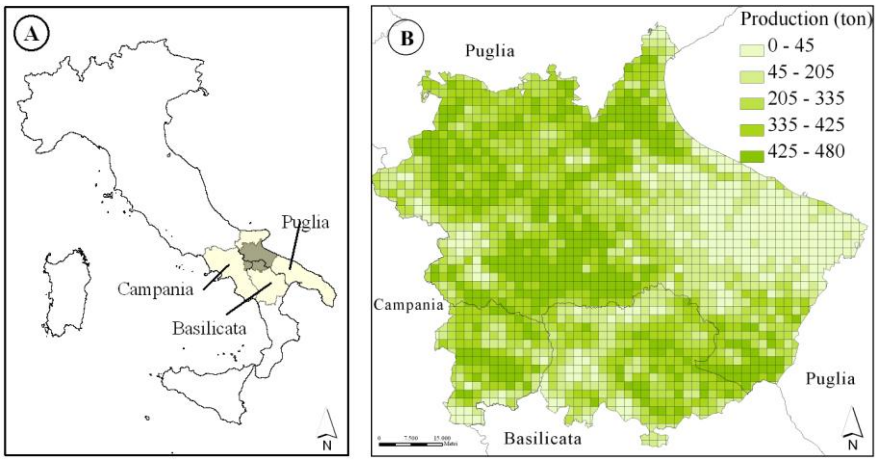


Figure 1. *A: area under study (in dark gray); B: map of the potentially available biomass from straws.*

Considering a hypothetical cell of the grid and supposing to locate the plant inside it, the distances D_i of each other cell from the chosen one can be determined. With respect to each cell, the corresponding F_i value can also be assigned. The cost estimation exactly followed the procedures reported by *Caputo et al., 2005*.
To test the effect exerted on *R*, *P* and *NPV* by different patterns of biomass spatial distribution, F_i has been changed from the actual values to those reconstructed in order to simulate three different conditions: *1.* the highest F_i values are assigned to clustered cells close to the plant (spatial decreasing biomass density); *2.* the highest F_i values are still assigned to clustered cells but far away from the plant (spatial increasing biomass density); *3.* constant F_i values (spatial uniform biomass density). These three different simulation scenarios were compared with each other. In a second set of simulations, the economic model was applied to the actual F_i but three different *Y* values (the reference

value, an increase and a decrease equal to 30%, respectively) were considered, the total available biomass being

$$Q_{tot} = \sum Q_i = S \cdot Y \cdot \sum F_i \, .$$

2.3 Spatial analysis. The plant location was determined applying a spatial "neighborhood" statistic function to the raster "biomass map". The statistic function is the "mean" and the neighborhood is a circular moving window. A "moving window" consists of a subset of the raster map; the result of the function is assigned to the central cell of the window, and the whole process is repeated for each cell in the map (*Varela et al., 2009*). In this study, the average biomass density of the grid parcels (t ha^{-1}) was considered. According to this procedure, spatial variation at the local level can be quantified and more details are revealed with a general smoothing effect on the original dataset (*Zhang et al., 2007*). In particular, as the window size used for calculation of neighborhood statistics increases, the smoothing effect of this statistical procedure became stronger, resulting in clearer patterns which emphasize the persistence of a certain number of areas with very high density values which can be eligible to plant location. For this purpose, six density maps were produced performing a neighborhood statistics according to a circular moving window whose radius varied from 2 to 7 cells (corresponding to 4 and 14 km).

3. Results

The "optimal" radius of the biomass supply basin, the one corresponding to the maximum NPV, showed to be independent of the particular spatial distribution of the F_i value within the grid (*Fig. 2.A*); this was invariably observed with respect to any of the three different F_i vector (decreasing, increasing and constant F_i values, respectively). Since the average F value was fixed and equal to 0.6 for the three vectors, P* and Q_{tot} are the same for the three simulations (*Fig. 2.A*). The "optimal" radius is also unaffected by the total available biomass Q_{tot} (*Fig.2.B*); clearly, an increased amount in the available biomass leaded to an increase in the value of P* and, consequently, in the maximum NPV. We can conclude that, at a certain R*, the higher is the value of Q_{tot} and the higher is the biomass density close to the plant location, the higher is also the NPV. *Fig. 3* shows the different spatial patterns which derive applying the neighbourhood statistics by increasing the radius of the moving window, from 2 to 7 cells. A progressive loss of details in spatial variation is associated to a simpler and clearer spatial biomass pattern. The smoothing effect tends to create larger homogeneous areas with lower

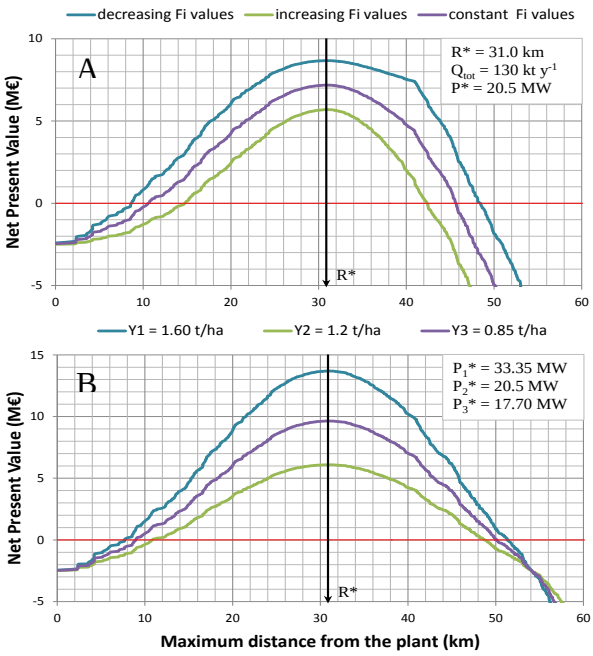


Figure 2. *Simulation results of the economic model. A: set of three different Fi vectors; B: set of three different Q_{tot} values.*

values of biomass density; nevertheless, the persistence of some cells with very high density values is still registered. The areas that are characterized by a persistently high biomass density, through different orders of scale, can be considered the most suitable for the location of the facility.

4. Concluding remarks

Results showed that the optimal radius of the supply basin was not related to (but independent of) both total amount and spatial distribution of biomass resources within the supply basin; differently, biomass availability strongly affected the size of the plant and the corresponding NPV. Therefore, the optimal plant location was at the center of the geographical area characterized by the highest biomass density.

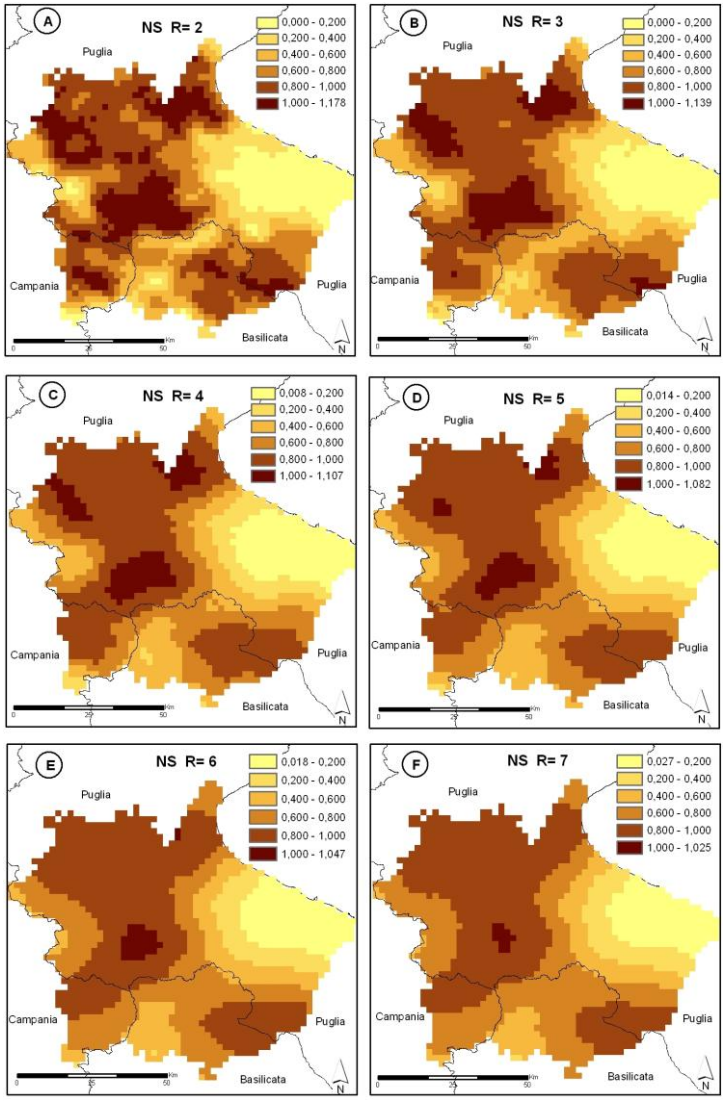


Figure 3. *Spatial distribution of the average biomass density values calculated using neighbourhood statistics (NS) with an increasing moving window radius (R from 2 to 7 cells).*

References

Alfonso D. et al. (2009) Methodology for optimization of distributed biomass resources evaluation, management and final energy use. *Biomass and Bioenergy*, 33, 1070-1079.

Caputo A.C. et al. (2005) Economics of biomass energy utilization in combustion and gasification plants: effects of logistic variables. *Biomass and Bioenergy*, 28, 35-51.

Leboreiro J. and A.K. Hilaly (2011) Biomass transportation model and optimum plant size for the production of ethanol. *Bioresource Technology*, 102, 2712-2723.

Rozakis et al., (2001) Multiple criteria decision-making assisted by GIS: Evaluation of Bio-Electricity Production in Farsala Plain, Greece. *Journal of Geographic Information and Decision Analysis*, (5)1, 49-64.

Varela et al. (2009) Multiscale delineation of landscape planning units based on spatial variation of land use patterns in Galicia, NW Spain. *Landscape Ecol. Eng.*, 5, 1-10.

Zhang C. et al. (2007) Using neighbourhood statistics and GIS to quantify and visualize spatial variation in geochemical variables: An example using Ni concentrations in the topsoils of Northern Ireland. *Geoderma*, 137, 466-476.

Population Density in a City

Corrado Abbate, Gianluigi Salvucci
Istat, abbate@istat.it; salvucci@istat.it

Abstract: The use of simple indicators may address towards incorrect assumptions about regions. As for cities, for example, population density could lead to wrong conclusions in social, economic and environmental analyses. We will demonstrate that the density of Italian cities' population makes Rome seem a rural city rather than a tertiary one, as it actually is. The aim of this paper is to show that some interpretations in socio-economic analysis are potentially wrong and to introduce some alternatives by using simple correctives like including environmental features. For instance, in the centre of Rome, where population is more concentrated, we have calculated a density of 55577 inhabitants per sq km versus the current estimation of 1981 inhabitants per sq km for the entire administrative territory.

Keywords: population density, spatial analysis, land cover.

1 Difficulties in describing a region by its density indicators

When data are related to different areas, a problem of comparability arises: it cannot be said that city A (1 million people) is bigger than B (5 thousand people) if area A is bigger than B. Statistics suggests to normalize the data by area for these cases. Urban geography has based a lot of its considerations on population density as an index for tertiary cities (Clark, 1951; Berry, Simmons and Tennant, 1963): the application of this indicator is supported by literature. Rome is the biggest Italian city, but is population density a real representation of the importance of a city, as suggested by literature? Rome is not in the top ten Italian cities by density.

Indeed, there is no significant correlation between a city's surface and its inhabitants. As a matter of fact, the R^2 index, calculated on the 8,101 Italian municipalities, is equal to 0.1465. It is clear that, if we want population density to mirror the importance of a city, we need to consider the variability in city areas' size. This is possible if one works with coarser resolution data, for example if one considers data aggregated by provinces.

If we think about cities, we imagine them as series of contiguous blocks, but reality is not always like that; we must introduce some information about land cover when interpreting population density. An useful suggestion is to divide a city into different zones and work separately on them. To do this, we need a very fine data resolution, for example referring to the smallest zones used in census cells (enumeration areas).

The EEA official CORINE Land Cover (CLC) dataset is suitable for our aims: it classifies the EU territory into land cover classes, with the aggregated class 1 meaning "Artificial surfaces" (including the sub-class 1.1.1 "continuous urban fabric"), class 2 meaning "Agricultural areas", 3 for "Forests and semi-natural areas", 4 and 5 for "Wetlands" and "Water bodies". By superimposing enumeration areas to a CLC dataset, by geographic coordinates of the perimeter of overlapping enumeration areas, we can obtain the predominant land cover class for each enumeration area and then aggregate

areas to re-calculate the density indicator (Table 1), where we can appreciate that it is not exhaustive to say that Italian population density is about 200 inhabitants per sq km, because the 79% of the population lives in just the 5% of the territory: the actual density in this 5% is 3563 inhabitants per sq km.

CLC level	Name	sq Km	Population	Density	% Area	% Populat.
1	Artificial fabric	12,262	43,694,310	3,563.4	5	79
2	Agricultural areas	135,575	10,817,936	79.8	60	19
3	Forests and semi-natural areas	77,715	1,090,776	14.0	34	2
	Total	225,552	55,603,022	246.5	100	100

Table 1 – Italian population distribution by CLC 2006 code and population density

Population rank	City	sq Km	Population	Density (Population/sqkm)	Density sq km for CLC class 1.1.1 .
1	Roma	1,285	2,546,804	1,981	14,682
2	Milano	182	1,256,211	6,899	16,152
3	Napoli	117	1,004,500	8,565	16,409
4	Torino	130	865,263	6,647	16,751
5	Palermo	158	686,722	4,322	14,892

Table 2 - Top five Italian cities by population, conventional density and density of 1.1.1 CLC class (our elaboration on CORINE land cover and Istat 2001)

Munic.	Land cover							
	CLC 1 pop.	*Percent*	CLC 2 pop.	*Percent*	CLC 3 pop.	*Percent*	CLC 4+5 pop.	*Percent*
Roma	2,390,042	*93.8*	155,904	*6.1*	786	*0.03*	72	*0.00*
Milano	1,241,329	*98.8*	14,847	*1.2*	19	*0.00*	16	*0.00*
Napoli	934,861	*93.1*	58,762	*5.9*	10,877	*1.08*	0	*0.00*
Torino	846,791	*97.9*	13,317	*1.5*	4,621	*0.53*	534	*0.06*
Palermo	669,076	*97.4*	12,762	*1.9*	4,884	*0.71*	0	*0.00*

Table 3 – Top five Italian cities' population distribution by city and land cover

Table 2 and 3 show different concentration for people living in different cities. Indeed, there is a bigger difference between Rome and Naples than between Rome and Milan.
Given all of this, we must consider that population does not live in industrial areas, so we should improve data resolution. These areas, like green areas, are not inhabited, and by including them we underestimate population density. In conclusion, only the 1.1 and 1.2 CLC classes (respectively, urban and industrial fabric) should be considered.
With 9409 people/sq km, the 1.1.1 class is the most densely inhabited type of land in Italy, while the 1.1.2 class (discontinuous urban fabric) is the most populated one only because it is the widest (7718 sq km versus the 1312 sq km of the 1.1.1 class). Population is more concentrated in 1.1.1 class, and density is the common criterion to define cities.

2 Some considerations about Rome

The core city is the area where population is more concentrated, and this is identifiable with land cover class 1.1.1, but we can consider the whole city area according to different land covers, to have different scenarios. Our analysis of Rome considers a radial city and the distribution of land cover and population.
Since Clark (1951), literature has studied the profile of cities without defining what a city and its centre are. When a city is not homogeneously populated, it is very difficult to get a good profile. To overcome this obstacle, we can use several variables related to urban profile, *i.e.* density, distance from mean weighted centre of population and land cover. We converted the CLC classification into 44 dichotomous variables: for each enumeration area a 1 value is assigned to the class corresponding to its prevalent land cover type , and a 0 value is assigned to the other 43 dichotomous variables. We decided to run a cluster analysis on these data to obtain a classification under a new urban perspective, considering all the variables at the same time (Table 4). Clusters may be constituted by different types of enumeration areas: cluster 4 is composed by continuous urban fabric and green urban areas, cluster 5 is composed by continuous urban fabric, discontinuous urban fabric and green urban areas.

		Prevalent Land cover code of enumeration area		
Cluster	Kms from centre	Continuous Urban Fabric (CLC 1.1.1)	Discontinuous Urban Fabric (CLC 1.1.2)	Green Urban Areas (CLC 1.4.1)
2	1.731	0	1	0
4	3.684	1	0	1
1	5.223	1	0	0
3	6.141	1	0	0
5	10.431	1	1	1

Table 4 - Clustering Rome enumeration areas by density, distance and land cover

Cluster	Km from centre	Area sq km	% Area sq km	Population	% Popul.	**Density**	N. of en. areas
2-4-1	3.546	7	0.5	374,031	14.7	**55,577**	818
3	6.141	48	3.7	1,120,155	44.0	**23,424**	3,560
5	10.431	1,233	95.8	1,052,618	41.3	**854**	8,721
Sum		1,287	100.0	2,546,804	100.0	**1,978**	13,099

Table 5 - Results of k-means cluster analysis for area and population

We found five clusters, but clusters 2 and 4 are very small, so we decided to join them to cluster 1. We obtain a partitioned-in-three city (Table 5): the first part includes the Central Business District, where land cover class 1.1.1 is predominant with the highest density (55577 people/sq km). The second cluster is less populated than the first one, with 23424 people/sq km and a prevalent land cover of class 1.1.1 too. The last cluster contains a mix of land covers, with a very low density compared to the other areas (only 854 people/sq km).
Considering all clusters, the analysis of urban profile should go from the city centre to the city boundary, and the central area could coincide (or not) with the most populated

zone. The presence of ruins in Rome city centre assigns this most densely populated area to "discontinuous urban fabric" 1.1.2 class.

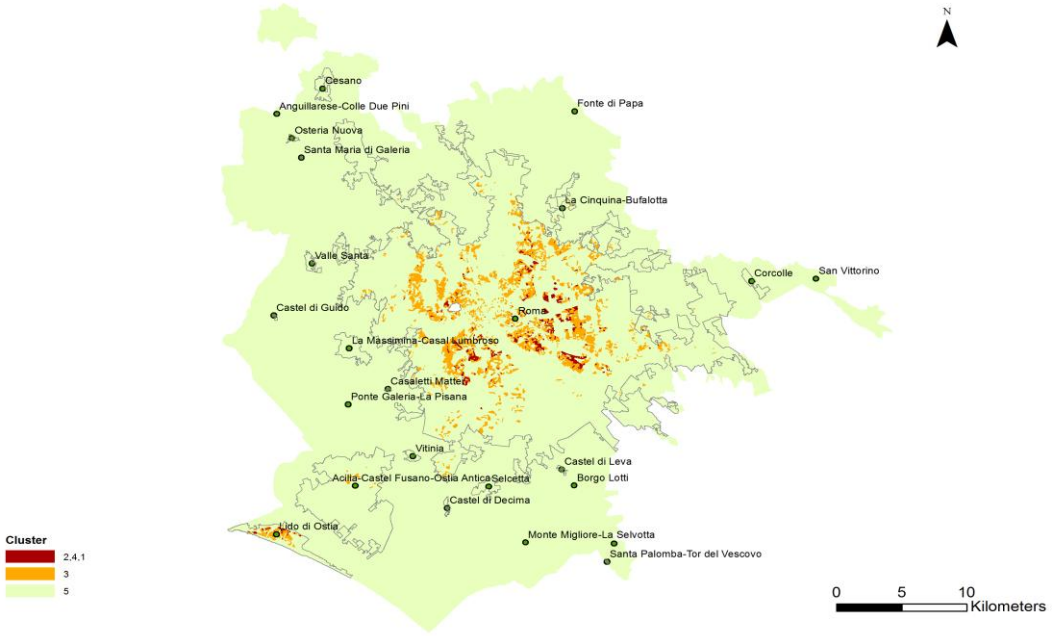


Figure 1 - Distribution of clusters in Rome

3 Conclusions

This paper aims at showing how social indicators must be integrated with environmental indicators to obtain a correct evaluation of a city population density.
Rome appears as a big green city with only 1981 people per sqkm, but, if we consider only the sub-areas where people actually live, we can evaluate a density of 55577 inhabitants per sq km in the most populated areas, and a density of 23424 people per sq km in the medium-populated ones; the 41.3% of the population lives in semi-agricultural areas with only 854 people per sq km (Figure 1).
In the city of Rome, the urbanized area hosts the 58.7% of the population, and it is undoubtedly very crowded. The simple density indicator does not allow a realistic evaluation of living conditions**.**

References

Berry, B., Simmons, J. and Tennant, R. (1963). Urban population densities: structure and change, *Geographical Review*, 12, pp. 389-405.

Clark, C. (1951). Urban population densities, *Journal of the Royal Statistical Society Series A* (general), pp. 490-496.

Comparison of spatial statistics for identifying underlying process in forest ecology

Calum Brown, Janine Illian
Centre for Research into Ecological and Environmental Modelling, University of St Andrews, calum@mcs.st-and.ac.uk

David Burslem
University of Aberdeen

Richard Law
University of York

Abstract: A number of different mechanisms have been suggested to explain species coexistence in diverse communities such as tropical rainforests. Spatial statistics appear to hold great potential for distinguishing the effects of these in empirical data, and a wide range of measures intended to describe spatial structure have been proposed. Using patterns generated by stochastic individual-based models, we examine the relative sensitivity of several of these measures to processes thought to be occurring in tropical rainforests, and so assess the potential for identifying specific coexistence mechanisms from empirical data. We then apply the measures to spatially explicit census data from a number of large-scale tropical rainforest plots in order to investigate the manifestation of ecological processes in forest spatial structure.

Keywords: spatial structure, coexistence mechanisms, tropical rainforest

1. Introduction

Statistics that summarise spatial pattern are of great interest in ecology, where a large number of processes influence, and are influenced by, spatial structure (Watt 1947; Bolker & Pacala 1997; Law *et al.* 2009). Spatial analysis is used for a wide range of purposes in plant ecology: for example to illuminate the relationship between environmental conditions and community structure (e.g. Kharuk *et al.* 2010; Obertegger *et al.* 2010); to study interactions between species (Hurlbert 1971; Wiegand 2007); and to isolate the signals of environmental and interactive effects and so assess their relative importance in producing observed community structure (Tuomisto *et* al 2003; Kraan *et al.* 2010). This is particularly important to attempts to investigate the processes that support the coexistence of species in diverse communities such as tropical rainforests (Brown *et al*. 2011). These processes may include niche differentiation, lottery dynamics, the Janzen-Connell effect, heteromyopia, or neutral drift.

The diversity of processes of interest has meant that a very large number of spatial summary statistics have been developed, even in place of those that have previously proved successful. These statistics tend to fall into discrete groups. Some of the most

established and widely-used deal with β-diversity (Whittaker 1972), summarising some aspect of the turnover in species composition with site. Measures of neighbourhood structure developed from spatial point process theory, however, represent the bulk of currently used spatial statistics (Wiegand & Moloney 2004; Illian *et al.* 2008).

While these measures have been useful both in descriptive and inferential studies of community ecology, their relative merits in detecting specific processes have been reviewed only infrequently (e.g. Koleff *et al*. 2003). In fact, many such measures share information used in their construction, and can be broken down into the individual counts or measurements which comprise them (Table 1). Furthermore, these can be considered in a multi-dimensional framework describing the level at which they operate. Information can be divided in this way between conspecific and heterospecific levels, scale-independent and scale-dependent, and individual, species or community level. The 'lowest' level information can therefore be seen as scale-independent descriptions of behaviour within species at the individual level; the 'highest' as scale-dependent multi-species community-level data. Measures of spatial structure use information from several different levels, often in combination, and can be formulated at higher levels by averaging some or all of the information they contain.

Here, we compare a limited number of popular measures of spatial structure on the basis of their ability to distinguish the spatial effects of models of species coexistence. Our aim is to determine which of the individual pieces of information which comprise these measures contain the most useful and robust signals. This allows for more accurate consideration of which information, and in what form, may best be used for the study of particular processes.

2. Materials and Methods

We consider a limited but representative number of measures of spatial structure that exemplify particular techniques for summarising spatial data. These measures can be divided between three broad groups – of β-diversity, within-species structure, and between-species structure. We consider three measures describing the spatial structure within species: the degree of aggregation; the measure of interspecific segregation; and the proportion of conspecific neighbours. All are intended to operate at the species or community level, although it is possible to calculate the proportion of conspecific neighbours at the individual level. Five measures describe spatial structure between species: the individual species-area relationship (ISAR); the mingling index; the spatial Simpson index; the degree of association; and the cross-pair overlap distribution (xPOD). Several measures of β-diversity are also included, and defined as by Koleff *et al.* (2003).

In order to test the sensitivity of these different measures to modelled ecological processes, we use data from stochastic individual-based models of a plant community which provide multispecies spatial patterns under neutral, niche, lottery, Janzen-Connell and heteromyopia assumptions. These were chosen as the principal theorised mechanisms of species coexistence in diverse plant communities.

n_j	number of individuals belonging to species j per unit area
n_k	number of individuals belonging to species k per unit area
$N_{jj}(r)$	number of conspecifics within a defined radius
$N_{ij}(r)$	number of heterospecifics within a defined radius
$N_{jk}(r)$	number of individuals belonging to species k within a defined radius
$N_{jk}(R)$	number of pairs of individuals belonging to species j and k separated by distance R (in practice, within range $(r + dr)$
A_c	area considered in count of points

Table 1: Separate pieces of information used in spatial measures considered here

3. Results

It is on the species level that most measures of spatial structure operate, making use of the numerous pieces of information which describe species-specific behaviour (Table 1). However, the differences these measures detect between models are often clearer when expressed at the community level. An example is the xPOD, which can be defined as:

$$A_{jk} = \int \log\left(\frac{N_{jk}(R)}{n_j n_k A_c}\right),$$

with terms as shown in Table 1. This measure describes the spatial overlap of all pairs of abundant species (with a threshold of 500 individuals) in a community, and shows substantial differences between models. Specifically, it shows that a far wider range of behaviour is produced by the niche and lottery models than any other, and the smallest range produced by the Janzen-Connell model. This suggests that species in the Janzen-Connell model are more mingled than under neutrality, and species in temporal or spatial niche models more segregated, on average.

These findings are confirmed by almost all of the other measures which we consider, and agree with theoretical predictions from each modelled process. Importantly, those measures which detect differences between the models all find higher levels of conspecific clumping and lower levels of heterospecific mingling in the niche and lottery models, and the opposite signals in the Janzen-Connell model. In addition, these signals are found in single pieces of information gathered at or averaged to the species level, prior to their combination to produce complete measures of spatial structure.

4. Concluding Remarks

In almost all measures (and at all levels), some aspect of the same behaviour is detected. In particular, the niche and lottery models produce clumped species which are not mingled, the neutral and heteromyopia models produce very similar spatial properties, and the Janzen-Connell model produces the least clumped and most mingled species. These findings are also apparent in single low-level pieces of information such as the proportion of conspecific neighbours, when expressed at the species or community level. In terms of β-diversity, those measures which emphasise simple counts of species unique to pairs of quadrats find the largest differences between models. This suggests both that the potential for distinguishing the modelled processes is limited to the spatial

characteristics listed above, and that relatively simple measures of spatial structure, operating at an appropriate level, have similar discriminatory power as those which are far more complex.

References

Bolker, B. & Pacala, S.W. (1997). Using moment equations to understand stochastically driven spatial pattern formation in ecological systems. Theoretical population biology, **52**, 179-197.

Brown, C., Law, R., Illian, J., Burslem, D. (2011). Linking ecological processes with spatial and non-spatial patterns in plant communities. *In review*.

Hurlbert, S.H. (1971). The nonconcept of species diversity: A critique and alternative parameters. Ecology, **52**, 577-586.

Illian, J.B., Penttinen, A., Stoyan, H. & Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*, 1st edn. Wiley, Chichester.

Kharuk, V.I., - Ranson, K.J., - Im, S.T. & - Vdovin, A.S. (- 2010). - *Spatial distribution and temporal dynamics of high-elevation forest stands in southern siberia*. - Blackwell Publishing Ltd.

Koleff, P., - Gaston, K.J. & - Lennon, J.J. (- 2003). - *Measuring beta diversity for presence?absence data*. - Blackwell Science Ltd.

Law, R., Illian, J.B., Burslem, D.F.R.P., Gratzer, G., Gunatilleke, C.V.S. & Gunatilleke, I.A.U.N. (2009). Ecological information from spatial patterns of plants: Insights from point process theory. Journal of ecology, **97**, 616-628.

Obertegger, U., - Thaler, B. & - Flaim, G. (2010). - *Rotifer species richness along an altitudinal gradient in the alps*. - Blackwell Publishing Ltd.

Tuomisto, H., Ruokolainen, K. & Yli-Halla, M. (2003). Dispersal, environment, and floristic variation of western amazonian forests. Science, **299**, 241-244.

Watt, A.S. (1947). Pattern and process in the plant community. Journal of Ecology, **35**, 1-22.

Wiegand, T., A. Moloney, K. (2004). Rings, circles, and null-models for point pattern analysis in ecology. *Oikos,* **104**(2), pp. 209-229.

Whittaker, R.H. (1972). Evolution and measurement of species diversity. Taxon, **21**, 213-251.

Wiegand, T., Gunatilleke, S. & Gunatilleke, S. (2007). Species associations in a hetergeneous sri lankan dipterocarp forest. The American naturalist, **170**, E77-E95.

Connectivity in a real fragmented landscape: distance vs movement model based approaches

Paola Mairota and Vincenzo Leronni
Department of Agriculture and Environmental Sciences
University of Bari "Aldo Moro" Italy
p.mairota@agr.uniba.it

Barbara Cafarelli
Department of Economical Mathematical and Statistical Sciences
University of Foggia Italy

Johannes Marinus Baveco
Alterra
Wageningen University and Research Centre The Netherlands

Abstract: Graph theory derived models and measures are increasingly being used to quantify landscape connectivity in order to contribute to conservation biology and management. This is particularly relevant in the case of real landscapes in which local actions may have crucial consequences for maintaining biodiversity on large scale. A number of graphs were compared sharing an identical node weight definition and whose link weights representing functional patch-connectivity, were derived from conceptually different approaches. Habitat suitability was taken into account. Calculated patch-connectivity was compared between all the graphs and these differences, evaluated by a set of indices describing network properties at the element structure level, were investigated.

Keywords: fragmentation, habitat suitability, matrix permeability, maximum entropy, graph theory, connectivity.

1. Introduction

Since the 1960's, the issue of species persistence in fragmented landscapes is crucial in both conservation biology and landscape ecology. Amongst other approaches, graph theory derived models and measures (Urban *et al.* 2009) are increasingly being used to quantify landscape functional connectivity in order to contribute to species and habitat conservation and management. Such tools have the potential to account for habitat availability, dispersal ability, species habitat requirements and dispersal route quality. These aspects are crucial to the conceptualisation and measurement of a landscape' permeability to the movement of organisms and thus to actually measure functional connectivity, as opposed to structural connectivity. However, landscape graph indices and models - as well as other techniques taking into account a heterogeneous landscape matrix - with desirable properties, may become too computation intensive for real large landscapes. The aim of this paper is to investigate the trade offs between a switch from binary landscape perspective to one embodying ecological continuity for a large real landscape.

2. Materials and Methods

The study area (EU NUT3 ITF45 Lecce, 275,716) is characterized by a very low forest share (1.4%) and a very high degree of fragmentation which challenge metapopulation dynamics (Hanski; 1991). One such dynamic is the dispersal of fleshy fruit broadleaved in pine plantations, likely to be mediated by bird species, among which the focal species was selected and described in terms of both breeding habitat and dispersal distance (5000 and 2500 m, 90-percentile). The habitat for the focal species was defined on two spatial data sets: 1) a 2008 land use vector map (1:5000 nominal scale) with potential breeding habitat (semi-natural woodland and plantations), and 2) a grid map (resolution 50 m) with probabilities of species-geographic distribution as a proxy to habitat suitability. These probabilities were obtained by applying an Environmental Niche Model (MaxEnt, Phillips and Dudík, 2008). The model was run using presence data (128 points) from a sub-regional ornithological monitoring program (La Gioia and Scebba, 2009). Several environmental predictor variables (i.e., land use, climate, landform, density of water elements and semi-natural vegetation), Linear Quadratic Hinge feature and a regularisation parameter equal to 3.0 , to compensate for potential overfitting, were considered in the model specification. The habitat system was cast in terms of graph theory, as a graph *G*, consisting of *n* nodes connected by *m* links. A node here is a functional unit: a patch with a local population, obtained from the clustering of nearby fragments likely to exchange individuals, within 250 m, which also served to greatly reduce the number of units, while preserving the exact habitat area. Patch population size is expressed as potential number of breeding pairs (reproductive units, RU) for which focal species is proposed as a measure of node weight (w_i). RU is determined by the area of suitable habitat and quality of the area. This is obtained by combining the definition of breeding habitat (vector format), with the MaxEnt derived definition of quality (raster format). Four graphs, two for each dispersal distance, were generated with identical nodes and node weights but different links. These were calculated either from Euclidean distance (*D*) assuming a negative-exponential relationship or with a simplification of the original GRIDWALK stochastic grid-based movement model . Distance-based links are symmetrical, as opposed to movement model based asymmetrical ones. The graph analysis was made as follows. Firstly, the weights of all links and the distance-based values (p_d) vs movement-based ones (p_m) were compared. Secondly, a set of published index, were based on the *PC* index routinely used for landscape conservation planning and change monitoring applications (Saura and Rubio 2010). These indices were compared at element level (Rayfield *et al.* 2011) by means of the measure of the individual patch's importance (*dPC*), and its breakdown into *dPC*(intra, flux, connector). The performance of a simplified, less computationally intensive, version of such indices was tested. In particular, *PCDP* and *DE* indices were considered. In PCDP index, the direct probabilities p_{ij},weighted by source and target node, are used instead of maximum product probabilities p^*_{ij} . The *DE* index (dispersal efficiency index),sums the values of all the fluxes in the graph. . In itsspecification a flux is defined as source node weight multiplied by link weight ($w_i \times p_{ij}$) and represents a relative measure of the number of dispersers expected to be exchanged between patches. For both indices we can define individual patch contributions, *dPCDP* and *dDE* as well. The map output similarities were evaluated by a fuzzy numerical approach (Hagen-Zanker *et al*., 2006, http://www.risks.nl/mck/), an extension to the numerical maps of Fuzzy Kappa method, generally used for comparing

categorical maps in order to account for fuzziness of locations and category. The comparison result is represented by a third map, indicating for each location the level of agreement in a range from 0 (non identical) to 1 (identical) between cells and by the similarity statistics evaluated as average of a combined one-way similarity over the whole map. An exponential decay function (2.5 km -5 km) was used for evaluating the similarities between maps in order account for the function used to evaluate the connectivity.

3. Results

The set of the statistical analysis on the model performance provided among MaxEnt model output information indicate a good model performance. As expected MaxEnt assigned different probabilities of distribution values to different patches (μ= 0.490, σ=0.184), and particularly to woodlands (μ= 0.672, σ=0.220) and plantation (μ= 0.553, σ=0.167) patches even though they belong to the same habitat type (i.e. suitable breeding habitat) for the focal species. This is because the model refers each focal habitat spatial element to its surrounding context conditions as defined by the niche factors fed into the model. Comparing distance-based with movement-based connectivity, we see little similarity. Differences were expected as the distance-based model ignores several factors that are known to affect the probability of encountering a patch, and that are taken into account in the movement-based values. A χ^2 test suggests complete independence between the variables. The distance-based values for the size of the target node (Moilanen and Nieminen 2002) were weighted by raising them to power of ½ in order to improve the correlation with the movement based ones. In general, the values of the distance-based approach are larger, providing a more optimistic view of connectivity. However, the impact of matrix heterogeneity is low: comparison of p_d with p_m values for a homogeneous matrix does not lead to a smaller χ^2 statistic. When directly comparing p_m for heterogeneous and homogeneous matrix the χ^2 values are very small, amounting to 0.0615 and 0.0867 for 2500 and 5000 m dispersal distance, respectively. For both the shorter and the longer dispersal distances considered (2500 m and 5000 m), the pairwise comparison shows a certain similarity between the *dPC* and *dPCDP* maps, as indicated by the values of similarity statistic which respectively assumes the values of 0.643 and 0.573. The similarity is weaker between *dPC* and *dDE* (0,410 and 0,480 respectively for the two distances). Indices *dPC_flux* and *dDE,* proxies for route specific fluxes, do not appear to be associated at neither distances (0.366 and 0.023).

4. Concluding remarks

It seems to be clear that by incorporating habitat quality (MaxEnt output) in the node weight, the resulting patch population carrying capacities were reduced in comparison to an approach based on the distribution of habitat only. However, the map defining matrix permeability, appeared to be relatively uniform at the local scale (50 m). As a consequence, we observed relatively little impact of matrix heterogeneity on connectivity, with p_m being relatively similar in homogeneous and heterogeneous landscapes. In this case, the value of working with a structured landscape matrix instead

of assuming a homogeneous matrix seems somewhat limited. This, far from contradicting the evidence that the matrix really does matter (Fisher *et al.*, 2008), indicates that the methods (including scale) we apply to estimate and express spatial heterogeneity, also matter. Distance and movement-based connectivity were very different but could be made more similar by correcting p_d with target patch size raised to ½. The extent to which correction is possible and it is however limited, as the real factor influencing accessibility (encounter rate) is the physical size of the patch accounting for shape as well, for which node weight (in *RU*) is just a weak approximation. In addition, there are several other factors determining accessibility in a movement-based approach, including 'shadowing' effects between patches, that are hard to correct for (but see . Likewise, it would be hard to correct for matrix heterogeneity. However, an interesting option appeared applying the movement model for a binary landscape. In this case, no assessment of landscape heterogeneity is needed, but still we implicitly deal with the impact of patch size and shape, and shadowing effects on patch connectivity. The large differences in underlying connectivity values (p_d versus p_m) do not translate into very different values of indices on the level of the nodes (*dPC* and *dPCDP*), the connected area metrics. We found a very high correlation between the index based on maximum product paths *dPC* and a comparable but simpler index based on direct probabilities *dPCDP*. Our results suggest that the latter may be used to substitute the first when dealing with large networks (>10^3 nodes and/or >10^5 links), reducing computation time from days to minutes. However, a more thorough analysis of the behaviour of *dPCDP* compared to that of *dPC* is required, to ensure that essential properties of *dPC* are preserved in the approximation.

References

Fischer J, Lindenmayer DB, Montague-Drake R (2008) The role of landscape texture in conservation biogeography: a case study on birds in south-eastern Australia. Diversity and Distributions 14:38–46

Hagen-Zanker A, Engelen G, Hurkens J, Vanhout R, Uljee I (2006) Map Comparison Kit 3: User Manual. Research Institute for Knowledge Systems, Maastricht

Hanski I (1991) Single species metapopulation dynamics: concepts, models and observations. Biological Journal of the Linnean Society 42:17-38

La Gioia G, Scebba S (2009) Atlante delle migrazioni in Puglia. Edizioni Publigrafic, Lecce pp 288

Moilanen A, Nieminen M (2002). Simple Connectivity Measures in Spatial Ecology. Ecology 83:1131-1145.

Phillips SJ, Dudík M (2008) Modeling of species distributions with MaxEnt: new extensions and a comprehensive evaluation. Ecography 31:161-175

Rayfield B, Fortin MJ, Fall A (2011) Connectivity for conservation: a framework to classify network measures. Ecology 92:847–858. [doi:10.1890/09-2190.1]

Saura S, Rubio L (2010) A common currency for the different ways in which patches and links can contribute to habitat availability and connectivity in the landscape. Ecography 33, 523-537.

Urban DL, Minor ES, Treml EA, Schick S (2009) Graph models of habitat mosaics. Ecology Letters 12: 260-273

Methodological study on pesticides in Alsatian groundwater

Fausta Musci, Concetta I. Giasi
Department of water engeneering and of Chemistry - Politecnico di Bari
Via Orabona 4, 70100 Bari, ITALY
f_musci@hotmail.com

Chantal de Fouquet
Centre de Géostatistique, Ecole des Mines de Paris, 77305, Fontainebleau, France

Abstract: The risk assessment conducted by many federal and state agencies have generally relied on deterministic approaches, that use single input/output values, generally selected to fulfill the goal of being health-protective. But, the presence of uncertainty and variability within the parameters of the procedure of risk assessment let them assume different values within a range of possible values, each with different probability of occurrence. In particular, the case study deals with groundwater contamination by agrochemical substances occurring in the French aquifer of Alsace. The regional supply of drinking water, water for irrigation and industrial water depends mainly on this water resource. A proper management of this area must consider, thus, the sustainability of a landscape capable of multiple uses and the overwhelming presence of censored data. For this reason, particular attention is given to the characterization of the extent and the chemico-physical distribution of the pollutant source for what concern the delimitation of the hazardous areas, to the determination of the probability density functions of the concerned variables and of the representative concentrations..

Keywords: groundwater contamination, geostatistics, estimation, non-linear methods.

1. Introduction

An instrument of high political and social importance is the risk assessment, or the evaluation of the risk associated with any event that can negatively affect the human health or the environment. Thus, the environmental impacts must be anticipated and prevented before they really happen and risk assessment has the logical structure to do it. The most immediate approach is therefore deterministic: by assigning to each of the input variables a single value, it gets a punctual value of risk. Every single value is generally selected to be reasonably certain that risk is not underestimated and to err on the side of overestimating risk. But, the presence of uncertainty and variability within the parameters of the procedure of risk assessment makes them actually random variables, as they are parameters that can assume different values within a range of possible values, each with different probability of occurrence. Therefore, these parameters can only be considered through a stochastic approach. In order to describe natural phenomena correlated in space and time and to quantify the uncertainty of the estimations of these phenomena carried on from a sampling generally very fragmentary, this work refers to the theory of the regionalized variables. It was developed by Matheron (1965) and then popularized by many others. In particular, the case study addresses the various method of linear and non-linear geostatistics for characterizing the exposure concentration through the inference of spatial structure from spot samples. Moreover, the overwhelming presence of censored data needs several statistical methods to be assessed, implemented and applied in order to characterize both variability and uncertainty of the exposure, effects and risk assessment.

2. Materials and Methods

2.1 Case study

In the Rhine valley, the alluvial formations create a large aquifer, one of the largest reservoirs of

drinking water in Europe. In the Alsatian part, this reservoir has the order of 45 billion m^3 of water for an area of 2800 km^2. The shallow depth of the groundwater makes its exploitation easy, which is an economic advantage. In fact, the groundwater provides three quarters of the drinking water needs of the population and more than half of the industrial and agricultural water needs. But besides this, the lack of protective geological cover and the shallowness of the aquifer make it particularly vulnerable to contamination due to human activities. And so, pesticides, as Atrazine, have been detected in the Alsatian groundwater.

Atrazine is an herbicide of the triazine chemical family, with radical absorption. It has been used in France on the cultures of corn since 1962, thus its use was prohibited by the 30 September 2003. Because it does not absorb strongly to soil particles (K_{oc} = 100 g/ml) and it has a lengthy soil half-life (60 to 100 days), it is expected to have a high potential for groundwater contamination, even though it is only moderately soluble in water (33 µg/ml). The Drinking Water Directive (DWD), Council Directive 98/83/EC, defines the sanitary thresholds (0.1 µg/l) for the concentration of these contaminants in drinking water.

The chosen data set is composed by four months of measurement: September 2002, March 2003, September 2003 and March 2004. This choice is based on the available samples (September 2003 is largely sampled – heterotopic case), the continuity of information available in time for each station, as well as the significance from the hydrological point of view. In fact, these months represent the beginning and the end of the recharge period of the aquifer. The period is also in correspondence of the interdiction of Atrazine's use in France.

2.2 Methodology of analysis

Geostatistics is based on the study of the spatial behaviour of variables. Even the concept of variable is converted in its spatial context as the regionalized variable [Matheron, 1965]. The model of the regionalized random variable is the basis principle of such kind of science.

The proposed procedure carries on through a series of steps, which will be deliberately presented in a synthetic and intuitive manner. For further details it is possible to refer to Matheron (1965, 1970), Chilès & Delfiner (1999), Rivoirard (1995) and Chauvet (1999).

1st. Exploratory data analysis. It refers to a statistical study of the data sets, for getting a first idea about data, their distribution, significance and consistency.

2nd. Structural analysis. It concerns all the methodologies aimed to investigate the spatial structure of data and exploit it to build reasonable spatial models. A synthetic form for explaining the structural variability of data is the experimental variogram. By fitting a continuous mathematical function on raw variogram it is possible to exploit such powerful instrument in order to model the variability structure for the whole spatial domain (not only on the measured points) [Isaaks and Srivastava, 1989].

3rd. Validation of a structural model. In practice it is important to evaluate the performance of fitting a variogram model.

4th. Local estimation. It allows passing from a discrete information to a continuous description of the phenomenon. The geostatistical estimator used for the estimation process is called kriging. For each target point, the linear estimator $Z^*(x_0)$ is expressed as the linear combination of the known points $Z(x_i)$, together with the conditions of unbiasedness and of minimization of the error variance.

5th. Multivariate aspects. Several regionalized variables could be treated together, so it is possible to enjoy also joint information that would increase the degree of accuracy of the results. The conjoint spatial structure of the variables is described by their cross-variogram and coregionalization models are used, between which the linear model of coregionalization [Journel & Huijbregts, 1978] is the simplest one. Thus the estimation is performed by cokriging.

6th. Non linear methods. This work refers to two principal methods: the probability from conditional expectation and the indicator cokriging. The first approach is of parametric type and is based on the "conditional expectation" estimator. It requires that the variable is multigaussian;

thus, first of all, a transformation of the original variable *Z(x)* – called anamorphosis ¬ is necessary for obtaining a random function *Y(x)* with gaussian distribution. It can be shown [Goovaerts, 1997] that the conditional distribution of *Y(x)* is Gaussian-shaped, with mean equal to its simple kriging $Y(x)^{SK}$ from the available data and variance equal to the simple kriging variance $\sigma^2_{SK}(x)$. Therefore, the posterior or conditional cumulative distribution function (in short, ccdf) at location x is

$$\forall\, y \in \mathbf{R},\ F(x; y \mid data) = G\left(\frac{y - Y(x)^{SK}}{\sigma_{SK}(x)}\right) \qquad [1]$$

where G(.) is the standard Gaussian cdf. Its implementation relies on an assumption of strict stationarity and knowledge of the prior mean m, in order to express $Y(x)^{SK}$ [Emery, 2006]. In the second approach, of non parametric type, the exceedance or not of a given threshold s (a concentration risk, for instance) at a point x can be coded by the indicator variable. Therefore, a cokriging estimation of the indicators could be performed, in order to consider simultaneously the indicator variables associated to different thresholds.

3. Results

Univariate linear geostatistical techniques have allowed obtaining estimation maps of Atrazine. This was a preliminary study of the data, which took account of the data set as measured, so without any kind of transformation, despite the highly asymmetric and discontinuous variables. In this case, the undefined values were exactly considered equal to their instrument detection limit (IDL). Same assumption has been made for the estimation in multivariate conditions, where variables are treated together, thanks to their significant correlations.

But kriging and its extensions provide what might be called, by abuse of language, the most probable value of the pollutant concentration at any point in space, combined with the variance of the error. This has two consequences. The first is that the map erases the "peaks" and "hollows" of pollution and is "attracted" by the average pollution on the area of interest: the real variability in the space of the pollution is not reproduced when the data are interpolated (smoothing property). The second consequence is that the complete distribution of the error is not accessible: just the mean (zero by construction) and the variance are known [Deraisme et al., 2003]. Therefore, these maps provide only an image more or less accurate of the reality. While, the comparison to a regulatory threshold needs to take into account the estimation error in order to reproduce the spatial variability. This is the object of non-linear methods. The proposed approaches reflect both the conditional expectation and indicator cokriging. While this second method can solve the uncertainty due to censored data, because all values are encoded in a binary variable [0,1], according to a certain threshold value bigger than the IDL, the first method is a bit more complex to implement. Needing a multi-Gaussian distribution, firstly it requires a parametric approach, performing thus a normalizing transformation of the strongly asymmetrical original data. These transformed variables must be thus multigaussian, that is to say every linear combination of the gaussian values should follow a gaussian distribution. In practice, the multigaussian hypothesis cannot be fully validated because, in general, the inference of multiple-point statistics is beyond reach. Usually, only the univariate and bivariate distributions are examined [Goovaerts, 1997]. However, the uncertainty of censored data persists, even in the Gaussian field, making the obtained transformed distribution inaccessible, and virtually impossible to analyze. The study then solves this "inaccessibility" through the use of indicator variables associated with different thresholds of the Gaussian transformed. In fact this allows, on the one hand, testing the bivariate normality (instead of the multivariate) of the obtained variables and, on the other hand, having a model to use in the estimation phase. Moreover, again because of this "inaccessibility" of the obtained transformation, the mean of the distribution is unknown, therefore, an approach via ordinary multigaussian kriging is preferred to using the simple kriging. In figure 1 the maps of atrazine obtained by linear estimation methods and non-

linear methods are reported, just for September 2003: tendentially they identify the same contaminated areas.

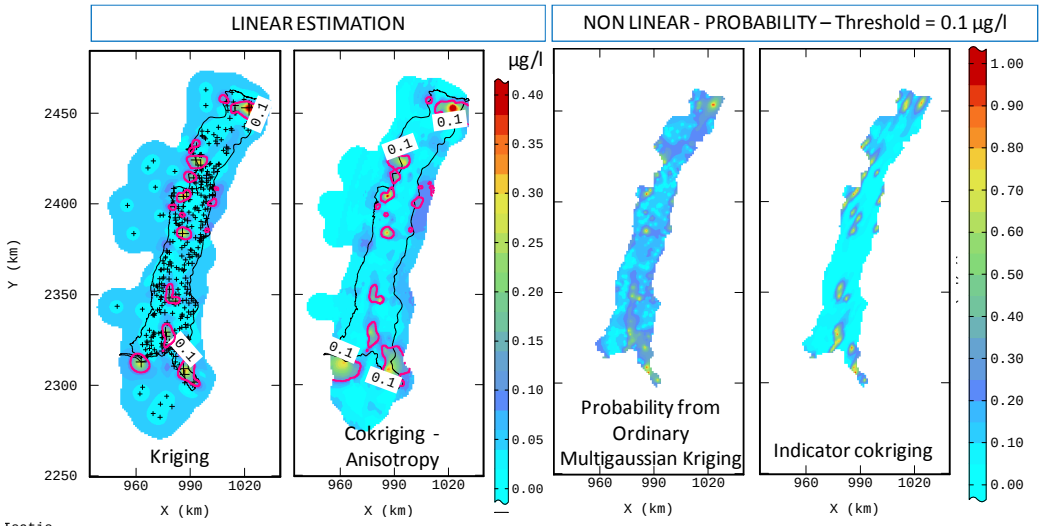


Fig. 1 Maps of atrazine obtained by estimation methods and non-linear methods – Sept 2003.

4. Concluding remarks

The proposed validation, through an ad hoc method of cross-validation, provided as result that the obtained probability by indicator cokriging is closer to the original data. The explanation for this result is that, probably, making a hypothesis of bivariate normality on the highly asymmetric and discrete available data sets is not unimportant in the estimation phase. Finally the risk results well characterized, also in function of the several assumptions and checks made during the analysis, and allows making considerations in terms of potential areas to be remediated and population potentially exposed to a hazard. Thus, the performed study is able to take into account uncertainty and variability related to the distribution of pesticides in groundwater in characterizing the scenario of contamination in the process of risk assessment. Moreover, the sensitivity analysis has allowed proceeding step by step in the study of the contamination by atrazine, considering limitations and advantages of the geostatistical methods, linear and non-linear. Finally, most of the methodologies presented in this study are also applicable in other field, as soil or air contamination.

References

Chauvet, P. (1999). «Aide-mémoire de gèostatistique linéaire». *Les Presses de l'Ecole des Mines*.

Chilès, J.P., Delfiner, P., (1999). «Geostatistics: Modeling Spatial Uncertainty». Wiley, New York.

Deraisme J., Bobbia M. (2003). « L'apport de la géostatistique à l'étude des risques liés à la pollution atmosphérique ». *Environnement, Risques & Santé* – vol.2, n°3, mai 2003.

Emery, X., (2006). «Ordinary multigaussian kriging for mapping conditional probabilities of soil properties». *Geoderma* 132 (1–2), 75–88.

Goovaerts, P., (1997). «Geostatistics for Natural Resources Evaluation». *Oxford University Press*, New York.

Isaaks, E.H. and Srivastava R.M. (1989). «An introduction to applied geostatistics». *Oxford University Press.*

Journel, A.G, Huijbregts, C.J. (1978). «Mining Geostatistics». *Academic Press*, London, 600 pp.

Matheron, G. (1965). «Les variables régionalisées et leur estimation: une application de la théorie des fonctions aléatoires aux sciences de la nature». Masson, Paris.

Rivoirard, J. (1995). «Concepts et méthodes de la géostatistique». *Cours C-158*, Centre de Géostatistique, Ecole des Mines de Paris.

The GIS approach to detect the influence of the fresh water inflows on the marine-coastal waters: the case of the Apulia Region (Italy) through standard monitoring data

A. Porfido, E. Barbone*, V. La Ghezza, G. Costantino, V. Perrino, N. Ungaro, M. Blonda.
ARPA – Agenzia Regionale per la Prevenzione e la Protezione dell'Ambiente, Corso Trieste 27, 70126 Bari, Italia.
*Corresponding author: e.barbone@arpa.puglia.it

Abstract:
The water quality in the marine coastal areas is affected by "natural" features (geo-morphology, hydrology) as well as by the "human" use of land. Understanding the linkage between water quality and river catchments is fundamental for the evaluation of different options in the coastal zone management. Monthly monitoring surveys were performed by ARPA Puglia during the January 2008-December 2009 period, both in the Adriatic and Ionian Seas. Spatial and temporal patterns of water chemical-physical parameters and the trophic index (TRIX) were investigated using the GIS approach for the evaluation of the influence of freshwaters inflows on the coastal area. The results indicate the effectiveness of standard monitoring activities in the water quality control and the usefulness of the GIS tool in order to detect the influence of the river's runoff.

Keywords: GIS, mixed effect modeling, TRIX, coastal water, river runoff

1. Introduction

The coastal zones are areas where natural processes (change in precipitation inputs, erosion, weathering of terrain materials) as well as anthropogenic influences (urban, industrial and agricultural activities) are concentrated (Focardi et al, 2009). Among the anthropogenic activities, are worth of notice the over-utilization of groundwater resources, the pollution and discharge of wastewaters into the sea. Particularly, the rising water demand from agriculture in Southern Italy, an area with a natural water resource scarcity, leads to the accumulation of nutrients in river basins. An overall quantitative estimate of nutrients loadings runoff to the Mediterranean sea is reported by Strobl et al. (2009). Mixing between inland and coastal waters represents a key process for biological productivity, with strong implications for the whole coastal system functioning and ultimately on the fishery. For this reason, the understanding of linkage between river catchment and water quality in the associated coastal zone is fundamental for evaluation of different options in the integrated coastal zone management. The purpose of this study is to investigate the influence of the freshwater inflows on the marine-coastal waters of the Apulia Region by mean of GIS approach. Standard water monitoring surveys were performed by ARPA Puglia during January 2008-December 2009 period along Apulian coasts. Spatial and temporal surface distribution of TRIX index (Vollenweider et al., 1998) was examined to evaluate the trophic status and surface quality of the coastal water and was studied in relation to the physical (salinity) and biological (Chlorophyll a) parameters.

2. Materials and Methods

The standard water monitoring surveys were carried out at 15 transects along the Apulian coast (Fig.1). For each transect three sampling points, at distances ranging from 100 m to 3000 m, were chosen. Five transects were located in front of rivers mouth. Monthly water sampling of salinity, chlorophyll *a*, dissolved oxygen, dissolved inorganic nitrogen (DIN) and total phosphorus (TP) was performed from January 2008 to December 2009. In order to describe the spatial and temporal trend of TRIX, salinity and chlorophyll *a* in Apulian region coastal waters, we computed two types of maps for these parameters through interpolation via kriging using a linear variogram with slope=1 and anisotropy=1 (Golden Software, 2002). Namely, Surface Water Maps (hereafter SWM; Fig. 2) are calculated by space (transects) and time (months) interpolation of data collected at surface water, while Depth Water Maps (DWM; Fig. 3) are calculated by space (sampling points) and time (months) interpolation of data collected along the water column. The distribution of TRIX values among coast type and season have been analyzed through a mixed effect models, taking account the spatial autocorrelation of data (transect as random factor) and different variance-covariance structure (transect x station x season). The fixed factors were: season (four levels: winter, spring, summer, autumn); coast type (two levels: river mouth, coast); station (three levels: ST1_500 m, ST2_1000 m; ST3_3000 m). The initial full model was: Y = season x coast type x station. The model was refined, in order to define the fixed part, by manual backward stepwise selection using maximum likelihood to remove not significant terms. The resulting model was validated to verify the normality of residual.

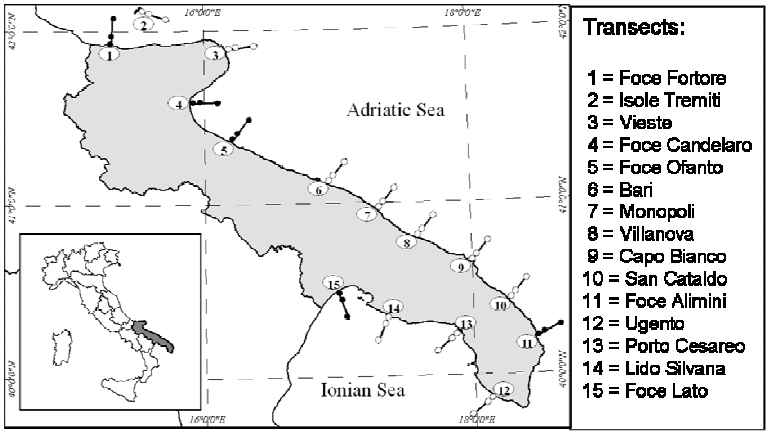


Figure 1: Study area and transects: river mouth stations (filled circles) and coastal stations (empty circles).

3. Results

SWM of salinity, chlorophyll *a* and TRIX value related to all transects were studied. An example is reported in Fig. 2. The river mouth transects showed different characteristics with respect to the coastal ones. In fact, mouth river transect was characterized by the presence of low salinity waters throughout the two years and high chlorophyll *a* and TRIX value (February-April 2008 and October-December 2009) with respect to the coastal ones. In order to investigate whether and how the river waters and coastal discharge affected the water column, the DWM of salinity and chlorophyll *a* were

studied along all transects. The layer of fresher waters and chlorophyll *a* were detected along the water column for the transects reported above (Fig.3). The selected statistical model indicates that TRIX index is affected by the interaction between coast type and 1) distance from the coastline 2) season (Fig. 4).

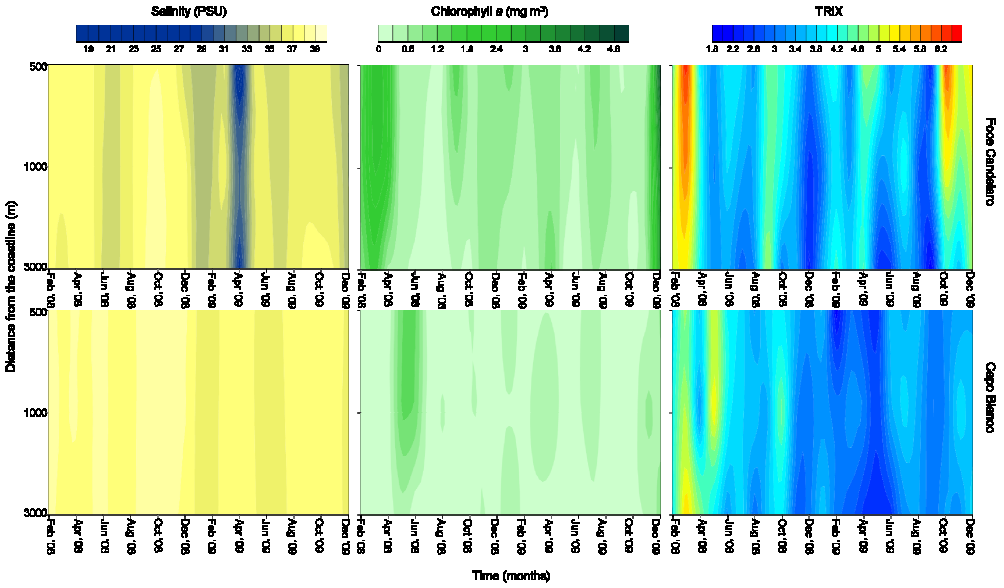


Figure 2: SWM of salinity, chlorophyll *a* and TRIX value in the river mouth (Foce Candelaro) and coastal transect (Capo Bianco) by kriging using a linear variogram with slope=1 and anisotropy=1. Note the not proportional distance among the stations.

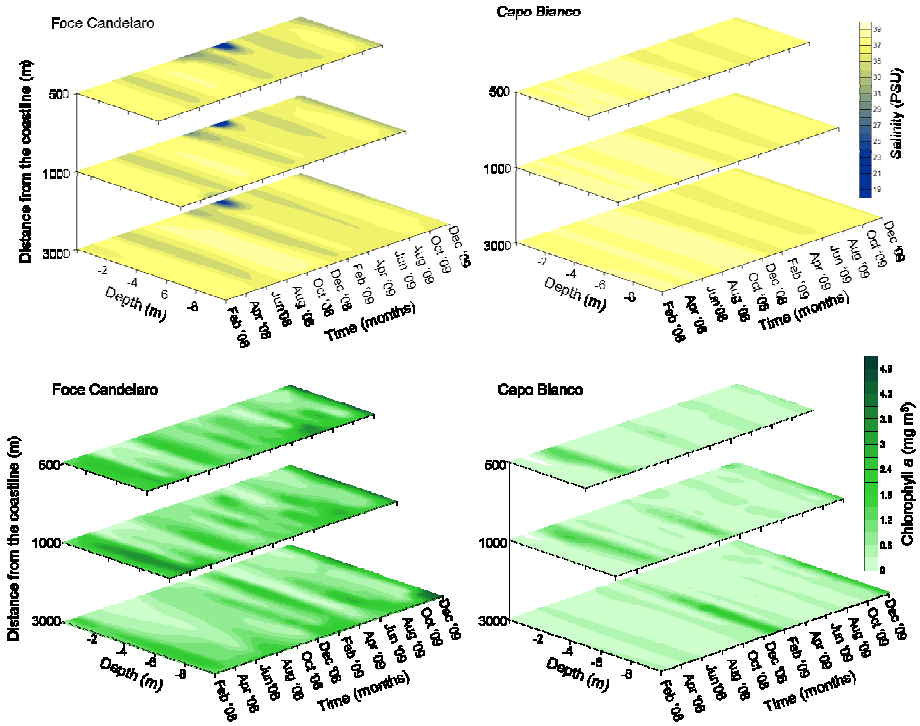


Figure 3: DWM of salinity and chlorophyll *a* in the river mouth (Foce Candelaro) and coastal transect (Capo Bianco) by kriging method using a linear variogram with slope=1 and anisotropy=1.

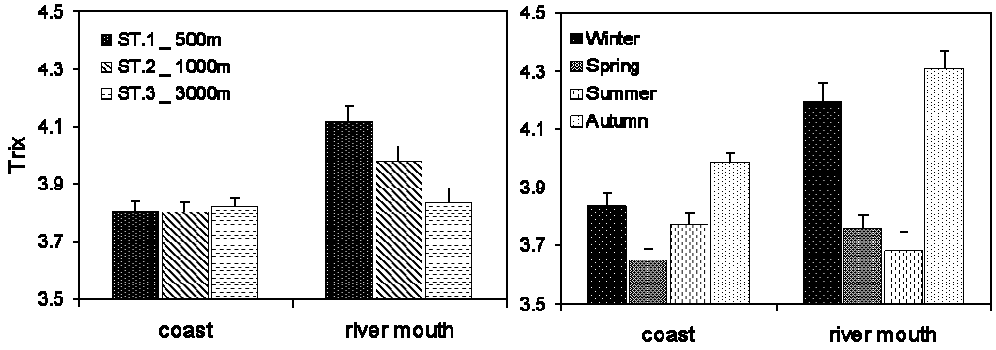


Figure 4: Mean values (± 1 e.s.) of TRIX index among coast type and i) distance from the coastline (left) ii) season (right) following the result of the final mixed linear model.

4. Concluding remarks

The investigation of the influence of freshwater on the marine-coastal waters was carried out at different sites of Apulia region, using the distribution of the TRIX index. With regard to the Northern Adriatic Sea, previous studies are available on the influence of salinity on the distribution of TRIX index on the coastal environment (Cocchi and Scagliarini, 2005). This work represented the first study regarding the role of land-derived water discharge in the coastal area in Apulia region, where most of the rivers are torrent-like characterized by temporary regime (Anonymous, 2005). The main results highlight the influence of the river runoff on the coastal water quality. A clear-cut distinction between rivers' mouth and coastal sites is revealed by means of statistical analysis. The highlighted differences are related to: low and homogenous TRIX values along coastal transects, while higher values and a gradient from inshore to offshore have been detected according to rivers' mouth sites.

References

Anonymous (2005). Piano di tutela delle acque della regione Puglia.

Cocchi D., Scagliarini M. (2005) Modelling the effect of salinity on the multivariate distribution of a water quality index. *Journal of Mathematics and Statistics*, 1 (4), 268-272.

Focardi S., Specchiulli A., Spagnoli F., Fiesoletti F., Rossi C. (2009) A combinated approach to investigate the biochemistry and hydrography of a shallow bay in the South Adriatic Sea: the Gulf of Manfredonia (Italy), *Environmental Moniitoring Assessment*, 153, 209-220.

Strobl R. O., Somma F., Evans B. M., Zalvìdar J. M. (2009) Fluxes of water and nutrients from river runoff to the Mediterranean Sea using GIS and a watershed model, *Journal of Geophysical Research*, Vol.114, G03012.

Surfer. 2002. Surface Mapping System, Surfer version 8.0. Golden Software, Inc.

Vollenweider R. A., Giovanardi F., Montanari G., Rinaldi A. (1998) Characterization of the trophic conditions of marine coastal waters with special reference to the NW Adriatic Sea: proposal for a trophic scale, turbidity and generalized water quality index. *Environmetrics*, 9, 329-357.

Applying a new procedure for fitting a multivariate space-time linear coregionalization model [1]

Sandra De Iaco
Dip.to di Scienze Economiche e Matematico-Statistiche, Facolta' di Economia, Universita' del Salento, Italy, sandra.deiaco@unisalento.it

Monica Palma
Dip.to di Scienze Economiche e Matematico-Statistiche, Facolta' di Economia, Universita' del Salento, Italy.

Donato Posa
Dip.to di Scienze Economiche e Matematico-Statistiche, Facolta' di Economia, Universita' del Salento, Italy.

Abstract: The near simultaneous diagonalization of the sample space-time matrix covariances or variograms makes the fitting procedure of a space-time linear coregionalization model (*ST-LCM*) easier. The method is illustrated by a case study involving data on three environmental variables measured at some monitoring stations of the Puglia region, Italy. It is shown that the near diagonalization works very well for this data set and the cross validation results show that the fitted matrix variogram is appropriate for the data.

Keywords: space-time linear coregionalization model, simultaneous diagonalization, environmental data.

1 Introduction

In this paper, the new fitting procedure of a *ST-LCM* (De Iaco et al., 2011) based on the generalized product-sum variogram model, is illustrated through an application to a multivariate space-time data set concerning three environmental variables. This method, based on the simultaneous diagonalization of the matrix variograms computed for several spatial-temporal lags, makes the identification of the parameters of the *ST-LCM* very simple and flexible.

[1]Supported by Fondazione Cassa di Risparmio di Puglia.

2 The case study

The data set consists of ozone, O_3 ($\mu g/m^3$), Temperature (°C) and Relative Humidity (%) daily maximum values, collected during June 2009 at some monitoring stations of the Puglia region, Italy (Fig.1). The space-time correlation structure of

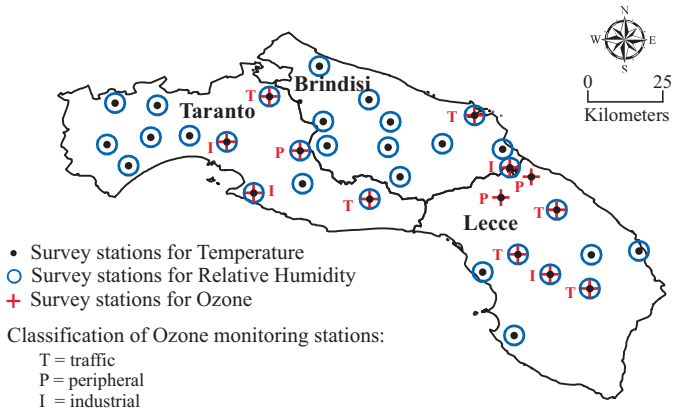


Figure 1: Posting map of survey stations in the South of Puglia region, Italy.

the variables under study has been modelled by a *ST-LCM*, whose basis components are generalized product-sum variograms.

2.1 Fitting process of a *ST-LCM*

The first step of the fitting process consists of computing the space-time direct and cross-variogram surfaces for the variables under study. Fig. 2 shows the variogram surfaces computed for 5 spatial lags and 10 temporal lags. Hence, 150 symmetric ($3\times$ 3) matrices of sample direct and cross-variograms have been obtained. Afterwards, the 150 symmetric (3×3) matrices of sample direct and cross-variograms have been simultaneously diagonalized by using the matlab code "joint_diag_r.m" (Cardoso, 1996). Hence, the orthogonal (3×3) matrix $\mathbf{\Psi}$ which simultaneously diagonalizes all these matrices is given below:

$$\mathbf{\Psi} = \begin{bmatrix} 0.9725 & -0.0719 & 0.2213 \\ 0.0969 & 0.9898 & -0.1041 \\ -0.2116 & 0.1227 & 0.9696 \end{bmatrix}. \quad (1)$$

Successively, by extracting the diagonal elements from the 150 diagonal matrices, the sample spatial-temporal variograms of the independent basic components have been obtained. Since the spatial and temporal marginal variograms of the second and third basic component show the same behavior, meaning that the spatial and temporal ranges are almost equal for the second and the third basic component, solely the first and the second basic component have been retained. Two different scales of spatial-temporal variability have been considered: 21 kilometers in space,

and 3 days in time, at the first scale of variability; 35 kilometers in space, and 8 days in time, at the second scale of variability. Hence, spatial and temporal marginal basic variograms, fitted to the empirical basic components have been the following:

$$\gamma_1(\mathbf{h}_s, 0) = 206\, Exp(||\mathbf{h}_s||/21), \qquad \gamma_1(\mathbf{0}, h_t) = 185\, Sph(|h_t|/3), \tag{2}$$

$$\gamma_2(\mathbf{h}_s, 0) = 3.1\, Sph(||\mathbf{h}_s||/35), \qquad \gamma_2(\mathbf{0}, h_t) = 8.3\, Exp(|h_t|/8), \tag{3}$$

where $Sph(\cdot)$ and $Exp(\cdot)$ are the abbreviated forms for the spherical and the exponential models, respectively. The contributions associated to the first and the

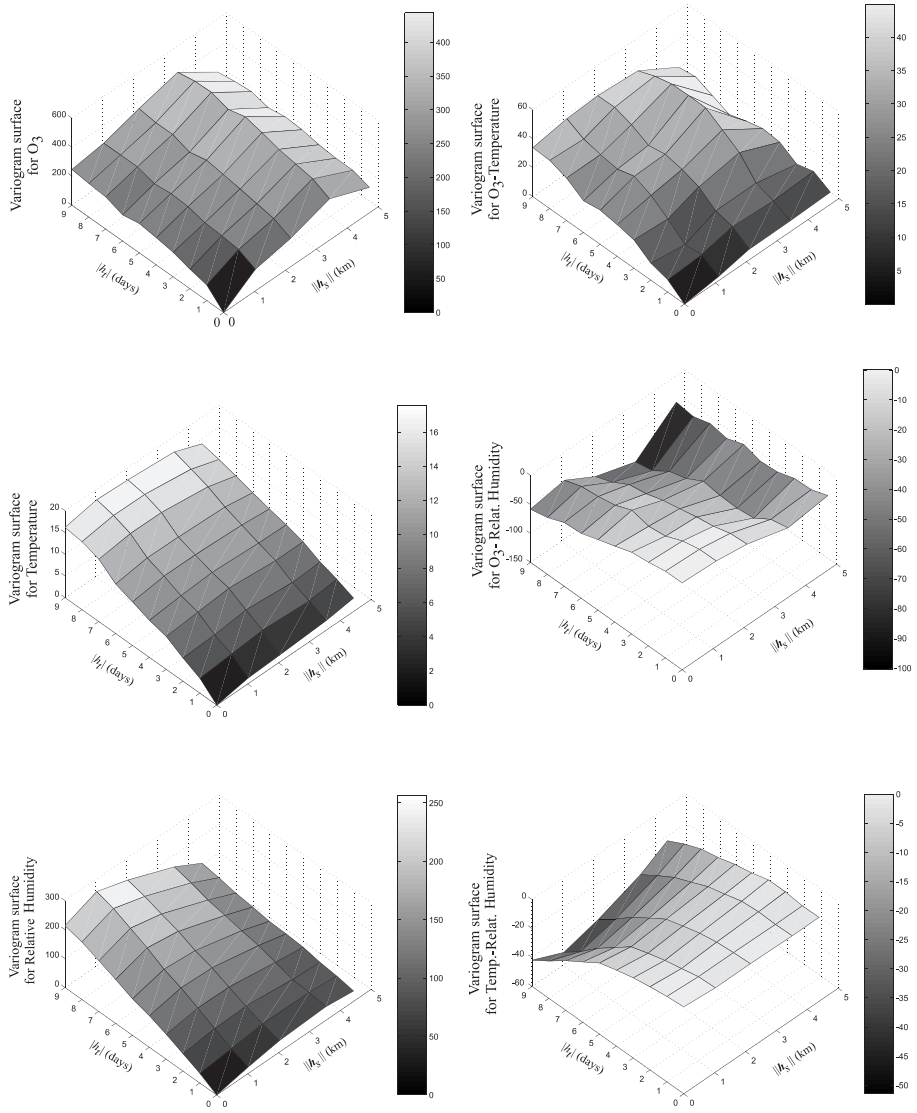


Figure 2: Space-time variogram surfaces of O_3, Temperature and Relative Humidity daily maximum values.

second basic components, i.e. the first and the second scales, are 293 and 8.7, respectively. Then, the coefficients $k_l, l = 1, 2$, are obtained as follows:

$$k_1 = \frac{206 + 185 - 293}{206 \cdot 185} \Rightarrow k_1 = 0.00257, \qquad k_2 = \frac{3.1 + 8.3 - 8.7}{3.1 \cdot 8.3} \Rightarrow k_2 = 0.10494.$$

Next, the entries in the matrices $\mathbf{B}_l, l = 1, 2$, have been computed:

$$\mathbf{B}_1 = \begin{bmatrix} 0.98635 & 0.08532 & -0.08360 \\ 0.08532 & 0.03038 & -0.01980 \\ -0.08360 & -0.01980 & 0.33447 \end{bmatrix}, \mathbf{B}_2 = \begin{bmatrix} 17.55814 & 2.09302 & -7.61628 \\ 2.09302 & 0.70930 & -2.45349 \\ -7.61628 & -2.45349 & 4.53488 \end{bmatrix}. \quad (4)$$

Hence, the *ST-LCM* for the analyzed variables is given below:

$$\mathbf{\Gamma}(\mathbf{h}_s, h_t) = \mathbf{B}_1\, g_1(\mathbf{h}_s, h_t) + \mathbf{B}_2\, g_2(\mathbf{h}_s, h_t), \quad (5)$$

where the matrices $\mathbf{B}_l, l = 1, 2$, are as in (4) and the space-time variograms $g_l(\mathbf{h}_s, h_t)$, $l = 1, 2$, are modelled as a generalized product-sum model as follows:

$$g_1(\mathbf{h}_s, h_t) = \gamma_1(\mathbf{h}_s, 0) + \gamma_1(\mathbf{0}, h_t) - k_1\gamma_1(\mathbf{h}_s, 0) \cdot \gamma_1(\mathbf{0}, h_t),$$
$$g_2(\mathbf{h}_s, h_t) = \gamma_2(\mathbf{h}_s, 0) + \gamma_2(\mathbf{0}, h_t) - k_2\gamma_2(\mathbf{h}_s, 0) \cdot \gamma_2(\mathbf{0}, h_t),$$

where $\gamma_1(\mathbf{h}_s, 0)$, $\gamma_2(\mathbf{h}_s, 0)$, and $\gamma_1(\mathbf{0}, h_t)$, $\gamma_2(\mathbf{0}, h_t)$, are the spatial and temporal marginal basic variogram models respectively, as indicated in (2) and (3), while k_1 and k_2 are the coefficients of the model.

2.2 Validation of the fitted coregionalization model

Using the modified *GSLib* routine "COK2ST" proposed in De Iaco et al. (2010), cross-validation has been performed in order to evaluate the goodness of the fitted *ST-LCM*. O_3, Temperature and Relative Humidity daily maximum values have been estimated at all data points by space-time cokriging using model (5). The proportion of absolute normalized errors (normalized by the cokriging standard deviations) exceeding 2.5, is very small (less than 1.5%) for each variable: from Chebyshev's inequality this proportion should be less than 1/6.25 (i.e, 16%). Hence, the fitted matrix variogram can be considered appropriate for the observed data.

References

Cardoso, J.F., Souloumiac, A. (1996) Jacobi angles for simultaneous diagonalization, *SIAM J. Mat. Anal. Appl.*, 17, 161-164.

De Iaco, S., Myers, D.E., Palma, M., Posa, D. (2010) FORTRAN programs for space-time multivariate modeling and prediction, *Comput. & Geosc.*, 36, 5, 636-646.

De Iaco, S., Palma, M., Posa, D. (2011) A new procedure for fitting a multivariate space-time linear coregionalization model, *Proceedings of 2011 European Regional Conference in Spatial Data Methods for Environmental and Ecological Processes, 1-2 September 2011, Foggia (Italy).*

Decision making for root disease control: a problem in reducing the nugget variance[1]

Ray Correll
Rho Environmetrics rho.environmetrics@bigpond.com

Margaret Evans, Robin Harding and Alan McKay
South Australia Research and Development Institute

Abstract: Root disease has the potential to cause large economic losses of agricultural production. Techniques are available for assessing the amount of pathogen present in soil using DNA assays. There is spatial variability in pathogen levels across fields and spatial methods would appear an obvious tool to use to map the incidence and distribution and as a basis to plan cropping programs. The information required for agronomic decisions has to be obtained in sufficient time and at an acceptable price for this to be a viable technique. Two examples where this is being used are wheat in large (40 to 100 ha) fields and potatoes grown in centre pivots. The largest difficulty encountered is obtaining a sampling scheme that produces a small nugget variance. Alternative sampling strategies are considered.

Keywords: Disease mapping, Agriculture and biodiversity

1. Introduction

Sampling fields for nutrient levels is used as a tool to optimize inputs and profits. More recently with the development of DNA based testing services, growers can now measure pathogen levels as an indicator of the disease potential of a field.

The challenges in sampling are to provide a 'fit for purpose' sampling scheme. Currently sampling is achieved by using cores of soil, or some alternative sampling scheme that provides a uniform representation of say the top 100 mm of soil. The sampling scheme has to be unbiased and it should have a small coefficient of variation and be cost efficient. The variation among samples arises from local variation (on a scale much less than 1 m) and also on a much larger scale (100 m or more). The local variation (nugget effect) represents variation between samples taken close together.

One method of reducing the local variation is to take many (up to 40) cores and combine them to form a composite sample. (Note that in some literature a core may be referred to as an aggregate sample and the composite as a cumulative sample). The

[1] Funding was provided by the Grains Research and Development Corporation Australia, South Australian Grain Industry Trust and the Australian Potato Research Program".

composite samples represent an average of many core samples and hence should be more representative (have smaller coefficient of variation) of the local area.

Commonly cores to be composited are taken in some convenient predetermined pattern – perhaps in a circle around a vehicle. Alternatively samples could be taken in a straight line. The composite sample ideally should represent the range of variation near the nominated sampling point. Conventional wisdom says that increasing the distance apart of samples typically increases the variation between samples. A convenient and effective pattern for local sampling is therefore to take a series of small cores along a straight line. This suggests an alternative method of sampling where the sample is taken as a slot that is cut using a circular saw. Such a device is known as a linear sampler. It effectively takes samples from a line simulating the effect of taking many cores in a straight line.

A known source of local variation is the crop rows. These could hold increased levels of pathogen DNA compared to the inter-row. Furthermore there could be differences in nutrient status as fertiliser is usually applied with the drill. Differences between the row and inter-row are therefore to be expected. An alternative approach is to sample across the rows – this can be simply achieved with the linear sampler.

A trial has been carried out at two sites to explore variation in wheat. Each composite sample was assessed for phosphate (as representative of nutrient status) and for *Fusarium pseudograminearum* a stubble borne pathogen typically concentrated in the row, which causes to assess pathogen status of the fields. Data on six pathogens (including Black Dot, *Colletotrichum coccodes*) have also been collected from a range of potato crops that are grown using centre pivot irrigation. These data have been obtained from composite samples each representing one ha, with 40 cores used for every composite.

2. Materials and Methods

Linear sampler: A linear sampler was used to take some of the samples in the wheat field. The linear sampler essentially consists of s circular saw mounted on a carriage. The saw cuts a 10 cm deep slot in the soil (Figure 2) and output collected. Care was taken to ensure that the soil is representative of the 0 – 10 cm soil layer.

Core sampling of wheat fields: The wheat field was sampled at 26 sites in a systematic pattern to represent the area. At each sampling site, 10 cores were taken on the stubble row, 10 between the stubble rows, 10 on the stubble rows, a linear sample between the rows and two separate linear samples cross the rows.

Sampling of potato soil: Potatoes were grown under a centre pivot, with each pivot covering 30 – 50 ha. A 100 m square grid (1 ha) was superimposed on each centre pivot and a single composite sample was formed from 40 cores (12 mm diameter and 100 mm deep) taken along a W shaped transect from each ha. Each composite sample was assessed for potato root pathogens.

Variograms were constructed using the 'geoR' package (Ribeiro Jr. & Diggle 2001) using the 'variofit' function with 'max' set to half the observed maximum distance.

3. Results

Wheat root disease

There was effectively no correlation between of the pathogen DNA data obtained by the different sampling methods with correlations of the phosphate data ranging from 0.44 to 0.70 for samples taken within two metres of each other.

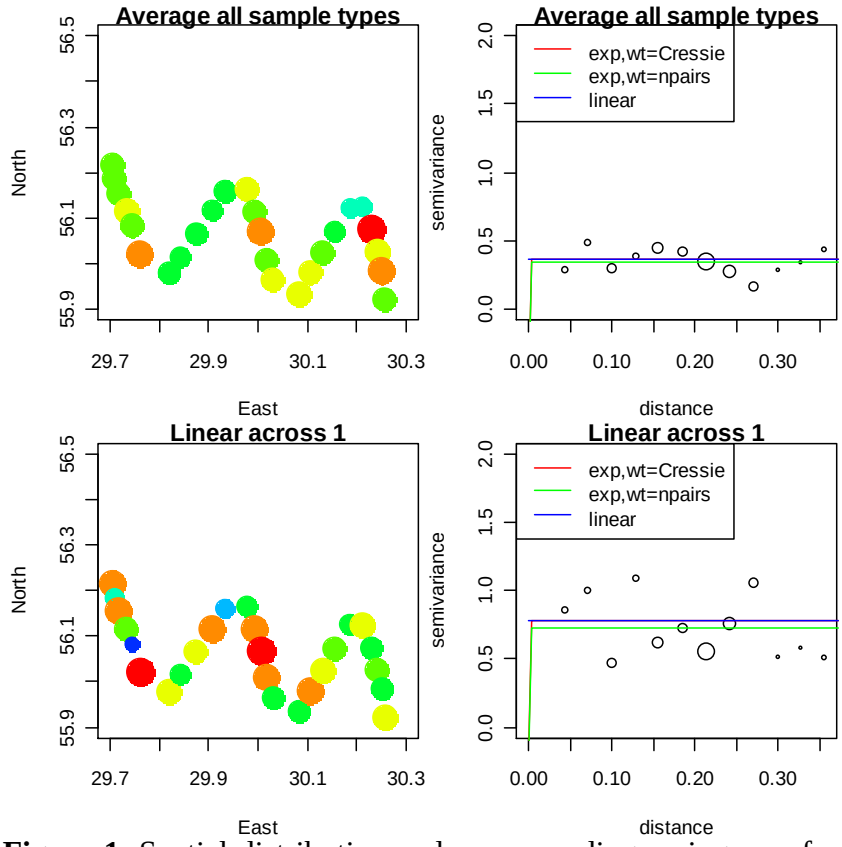


Figure 1: Spatial distribution and corresponding variograms for crown rot at Kybunga for average over a sample types series of samples of root disease (top half) and results from the linear sampler (bottom half). High amounts of disease are indicated by large red circles grading to low amount of disease shown as small blue circles. Each unit of distance is approximately 1 km.

Figure 1 shows the spatial distribution and estimated variogram for one series of linear samples. In most cases the variograms were not stable and the exponential fits would not converge for the inter-row samples (data not shown). Two sampling schemes (cores between rows and linear sampler on the row) showed a large nugget effect but also an increase in variance with distance. Although the data were expressed on a natural log scale, there were high estimates of the variance even when the samples were close together. This was despite using composite sampling or the linear sampler. Even when all six sampling types were averaged there was still no evidence of increasing variance with distance, but the estimated variance approximately halved (Figure 1).

Potato root disease

Potato disease shows some spatial patterning (but the sampling c.v. is still well over 50%). The commercial reality is that a farmer will use a maximum of four samples to represent a pivot of approximately 40 ha when making management decisions. Data from each ha are available for research purposes and these have been used to give a good approximation of the proportion of area that has pathogens levels that exceed an acceptable level. Detailed data are available to assess how well 4 samples can represent a pivot. A simulated commercial sampling was achieved by choosing a single ha from each quadrant of a pivot and obtaining its mean. The proportion of the pivot that exceeded a risk level was plotted against the mean DNA level of the four samples (Figure 2 right panel) and a logistic distribution was fitted. The results indicated that data from each of four 1 ha samples can be used to give a 'correct' answer in about 85% of cases despite the variability of the sampling.

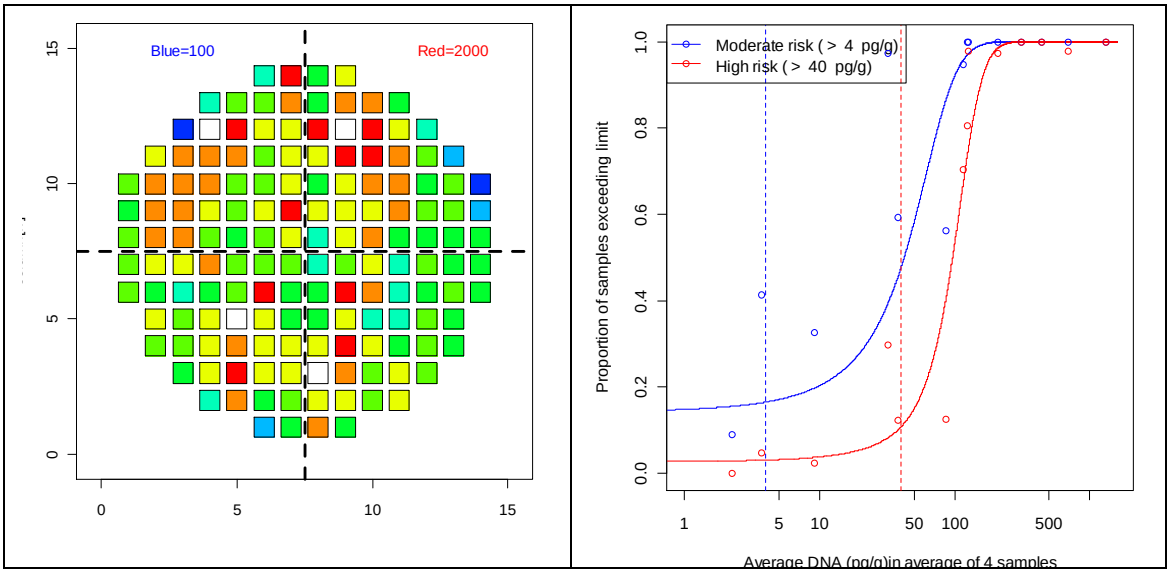


Figure 2: Left panel shows distribution of BD DNA in a typical pivot. Right panel indicates proportion of correct decision would have been made based on 4 samples. High and low levels indicate currently recognised limits of risk of commercial harm.

4. Concluding remarks

The number of samples available in a commercial agricultural application of the distribution of soilborne pathogens is far less than the number available for conventional spatial statistics. Furthermore, each sample has a very large variability (c.v. >50%). Despite these shortcomings, useful commercial decisions are currently being made.

The ongoing challenge to agricultural statisticians is how to take cost effective samples for the evaluation of root disease risk assessment, and to use our knowledge of spatial statistics to optimize this process. Nugget variance is a limiting use of spatial methods.

Reference

Ribeiro Jr., P.J. & Diggle .J. (2001) geoR: A package for geostatistical analysis. *R-NEWS*, **1** (2), 15-18.

EM estimation of the Dynamic Coregionalization Model with varying coefficients[1]

Francesco Finazzi and Alessandro Fassò
University of Bergamo - Dept. IIMM, alessandro.fasso@unibg.it

Abstract: The satellites from NASA's Earth Science Project Division, like AURA, produce data for the concentration of various airborne pollutants. Calibrating satellite data using ground level monitoring networks and other meteorological and land characterizing variables is mandatory. To do this, it is important to use an approach which is able to manage large datasets coming from different sources, structural missingness and spatial and temporal correlation. In this paper, we extend the Dynamic Coregionalization Model introduced in Fassò and Finazzi (2011) to the case of space-time varying coefficients in order to increase the model flexibility and to make it suitable for large regions such as Europe.

Keywords: air quality monitoring, missing data, dynamic coregionalization

1 Introduction

The Dynamic Coregionalization Model (DCM) of Fassò and Finazzi (2011) has been proven to be quite appropriate for modeling multivariate space-time environmental data in the non-collocated case and in the presence of missing data. When data are collected over continent-size regions, the statistical model considered must be enough flexible to accommodate for local conditions. In order to gain this flexibility, the DCM is extended here to the case of varying coefficients. The model is described in Section 2 and its estimation is addressed in Section 3.

2 The varying coefficients model

Let $\mathbf{y}(\mathbf{s},t) = (y_1(\mathbf{s},t),...,y_q(\mathbf{s},t))$ be the $q-$variate response variable at site $\mathbf{s} \in \mathcal{D} \subset \mathcal{R}^2$ and time $t \in N^+$. The model equation is

$$\mathbf{y}(\mathbf{s},t) = \mathbf{X}(\mathbf{s},t) \cdot \left[\mathbf{K}_{\mathbf{x}}\beta + \mathbf{K}_{\mathbf{z}}\mathbf{z}(t) + \sum_{j=1}^{c} \gamma_j \mathbf{K}_{\mathbf{w}}^{j} \mathbf{w}^{j}(\mathbf{s},t) \right] + \varepsilon(\mathbf{s},t) \qquad (1)$$

[1]This research is part of Project EN17, 'Methods for the integration of different renewable energy sources and impact monitoring with satellite data', Lombardy Region under 'Frame Agreement 2009'

where $\mathbf{X}(\mathbf{s},t)$ is a matrix of known coefficients (for instance $\mathbf{X}(\mathbf{s},t) = \mathbf{I}_q \otimes \mathbf{x}(\mathbf{s},t)$ is the $q \times (bq)$ diagonal block matrix built from the $1 \times b$ covariate vector $\mathbf{x}(\mathbf{s},t)$), $\mathbf{z}(t)$ is a latent $p-$dimensional temporal state with markovian dynamics $\mathbf{z}(t) = \mathbf{G}\mathbf{z}(t-1) + \eta(t)$ with $\mathbf{G}$ a stable transition matrix and $\eta \sim N(0, \mathbf{\Sigma}_\eta)$ while each $\mathbf{w}^j(\mathbf{s},t) = \left(w_1^j(\mathbf{s},t), ..., w_q^j(\mathbf{s},t)\right)$, $1 \leq j \leq c$ is a q-dimensional gaussian latent coregionalization component with covariance and cross-covariance matrix function $\mathbf{\Gamma}_j = cov\left(w_i^j(\mathbf{s},t), w_{i'}^j(\mathbf{s}',t)\right) = \mathbf{V}_j \rho_j(h,\theta_j)$, $1 \leq i, i' \leq q$, $1 \leq j \leq c$. Each $\mathbf{V}_j$ is a correlation matrix and each ρ_j is a valid correlation function parametrized by θ_j. Finally, $\varepsilon(\mathbf{s},t) = (\varepsilon_1(\mathbf{s},t), ..., \varepsilon_q(\mathbf{s},t))$ is the measurement error which is assumed to be white-noise in space and time with $\varepsilon_i(\mathbf{s},t) \sim N(0, \sigma^2_{\varepsilon,i})$, $1 \leq i \leq q$.

The matrices $\mathbf{K_x}$, $\mathbf{K_z}$ and $\mathbf{K_w^j}$ are matrices of known coefficients which guarantee conformability of the model equation (1) and acts as selection matrices with respect to the columns of $\mathbf{X}(\mathbf{s},t)$. The model parameter set is $\Psi = \{\beta, \sigma^2_\varepsilon, \mathbf{G}, \mathbf{\Sigma}_\eta, \gamma, \mathbf{V}, \theta\}$ where $\beta = (\beta_1, ..., \beta_q)'$, $\sigma^2_\varepsilon = \left\{\sigma^2_{\varepsilon,1}, ..., \sigma^2_{\varepsilon,q}\right\}$, $\gamma = \{\gamma_j, ..., \gamma_c\}$, $\theta = \{\theta_1, ..., \theta_c\}$ and $\mathbf{V} = \{\mathbf{V}_1, ..., \mathbf{V}_c\}$.

3 Likelihood function and missing data

At each time t, each variable y_i is observed over the set of spatial sites $\mathcal{S}_i = \{\mathbf{s}_{i,1}, ..., \mathbf{s}_{i,n_i}\}$, $1 \leq i \leq q$. The sets in $\mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_q\}$ are not constrained and can be disjoint. The observed vector at time t is then $\mathbf{y}_t(\mathcal{S}) = (\mathbf{y}_{1,t}(\mathcal{S}_1), ..., \mathbf{y}_{q,t}(\mathcal{S}_1))' = \mathbf{y}_t$ and it has dimension $N = n_1 + ... + n_q$. The observation equation is $\mathbf{y}_t = \mu_t + \varepsilon_t$, where $\mu_t = \mathbf{U}_{\mathbf{x},t}\beta + \mathbf{U}_{\mathbf{z},t}\mathbf{z}_t + \gamma_1 \mathbf{U}^1_{\mathbf{w},t}\mathbf{w}^1_t + ... + \gamma_c \mathbf{U}^c_{\mathbf{w},t}\mathbf{w}^c_t$, $\mathbf{U}_{\mathbf{x},t} = \mathbf{X}_t\mathbf{K_x}$, $\mathbf{U}_{\mathbf{z},t} = \mathbf{X}_t\mathbf{K_z}$ and $\mathbf{U}^j_{\mathbf{w},t} = \mathbf{X}_t\mathbf{K}^j_{\mathbf{w}}$.

In the definition of the likelihood function, the distributions involved are

$$\begin{aligned}
\left(\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{w}^1_t, ..., \mathbf{w}^c_t\right) &\sim N_N\left(\mu_t, \mathbf{\Sigma}_\varepsilon\right) \\
\left(\mathbf{z}_t \mid \mathbf{z}_{t-1}\right) &\sim N_p\left(\mathbf{G}\mathbf{z}_{t-1}, \mathbf{\Sigma}_\eta\right) \\
\mathbf{w}^j_t &\sim N_N\left(0, \mathbf{\Sigma}^j\right), 1 \leq j \leq c
\end{aligned}$$

Let $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_T)$, $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_T)$ and $\mathbf{W}^j = \left(\mathbf{w}^j_1, ...\mathbf{w}^j_T\right)$. The complete-data log-likelihood function is given by:

$$-2l\left(\Psi; \mathbf{Y}, \mathbf{Z}, \mathbf{W}^1, ..., \mathbf{W}^c\right) = T\log\left|\mathbf{\Sigma}_\varepsilon\right| + \sum_{t=1}^{T}\left(\mathbf{y}_t - \mu_t\right)'\mathbf{\Sigma}_\varepsilon^{-1}\left(\mathbf{y}_t - \mu_t\right) +$$

$$T\log\left|\mathbf{\Sigma}_\eta\right| + \sum_{t=1}^{T}\left(\mathbf{z}_t - \mathbf{G}\mathbf{z}_{t-1}\right)'\mathbf{\Sigma}_\eta^{-1}\left(\mathbf{z}_t - \mathbf{G}\mathbf{z}_{t-1}\right) + \sum_{j=1}^{c} T\log\left|\mathbf{\Sigma}^j\right| \sum_{t=1}^{T}\left(\mathbf{w}^j_t\right)'\left(\mathbf{\Sigma}^j\right)^{-1}\mathbf{w}^j_t$$

At each time t, the observation vector $\mathbf{y}_t$ can be partitioned in the following way: $\mathbf{y}^*_t = \left[\begin{array}{cc} \mathbf{y}^{(1)}_t & \mathbf{y}^{(2)}_t \end{array}\right]'$ where $\mathbf{y}^{(1)}_t = \mathbf{L}_t\mathbf{y}_t$ is the sub-vector of the non-missing data and $\mathbf{L}_t$ is the selection matrix of the observed data at time t. The vector $\mathbf{y}^*_t$ is a

permutation of $\mathbf{y}_t$ and $\mathbf{y}_t = \mathbf{D}_t \cdot \left[\begin{array}{cc} \mathbf{y}_t^{(1)} & \mathbf{y}_t^{(2)} \end{array} \right]'$, where $\mathbf{D}_t$ is a permutation matrix. The partitioned measurement equation becomes $\mathbf{y}_t^{(l)} = \mu_t^{(l)} + \varepsilon_t^{(l)}$, $l = 1, 2$. and the variance-covariance matrix of the permuted errors is conformably partitioned, namely $Var\left[\left(\varepsilon_t^{(1)}, \varepsilon_t^{(2)}\right)'\right] = \left[\begin{array}{cc} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{array} \right]$. In what follows, $\mathbf{Y}^{(1)} = \left(\mathbf{y}_1^{(1)}, ..., \mathbf{y}_T^{(1)}\right)$ is the collection of the observed data.

4 EM estimation

At the E-step of the EM algorithm, the following conditional expectation is evaluated:

$$
\begin{aligned}
Q\left(\Psi, \Psi^{(k)}\right) &= E_{\Psi^{(k)}}\left[-2l\left(\Psi; \mathbf{Y}, \mathbf{Z}, \mathbf{W}^1, ..., \mathbf{W}^c\right) \mid \mathbf{Y}^{(1)}\right] \\
&= E_{\Psi^{(k)}}\left[E_{\Psi^{(k)}}\left[-2l\left(\Psi; \mathbf{Y}, \mathbf{Z}, \mathbf{W}^1, ..., \mathbf{W}^c\right) \mid \mathbf{Y}^{(1)}, \mathbf{Z}, \mathbf{W}^1, ..., \mathbf{W}^c\right] \mid \mathbf{Y}^{(1)}\right] \\
&= T \log\left|\mathbf{\Sigma}_\varepsilon\right| + tr\left(\mathbf{\Sigma}_\varepsilon^{-1} \sum_{t=1}^{T} \mathbf{\Omega}_t\right) + \\
&\quad T \log\left|\mathbf{\Sigma}_\eta\right| + tr\left\{\mathbf{\Sigma}_\eta^{-1}\left(\mathbf{S}_{11} - \mathbf{S}_{10}\mathbf{G}' - \mathbf{G}\mathbf{S}_{10}' + \mathbf{G}\mathbf{S}_{00}\mathbf{G}'\right)\right\} + \\
&\quad \sum_{j=1}^{c} T \log\left|\mathbf{\Sigma}^j\right| \cdot tr\left\{\left(\mathbf{\Sigma}^j\right)^{-1} \sum_{t=1}^{T} \mathbf{w}_t^{j,T} \cdot \left(\mathbf{w}_t^{j,T}\right)' + \mathbf{A}_t^{j,T}\right\}
\end{aligned}
$$

where:

$$
\begin{aligned}
\mathbf{\Omega}_t &= E_{\Psi^{(k)}}\left[\mathbf{e}_t \cdot \mathbf{e}_t' + \mathbf{\Lambda}_t \mid \mathbf{Y}^{(1)}\right] = E_{\Psi^{(k)}}\left[\mathbf{e}_t \cdot \mathbf{e}_t' \mid \mathbf{Y}^{(1)}\right] \\
&= \mathbf{D}_t \left[\begin{array}{cc} \mathbf{\Omega}_t^{(11)} & \mathbf{\Omega}_t^{(11)}\mathbf{R}_{11}^{-1}\mathbf{R}_{21} \\ \mathbf{R}_{21}\mathbf{R}_{11}^{-1} & \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{\Omega}_t^{(11)}\mathbf{R}_{11}^{-1}\mathbf{R}_{21} + \left(\mathbf{R}_{22} - \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\right) \end{array} \right] \mathbf{D}_t' \\
\mathbf{e}_t &= E_{\Psi^{(k)}}\left[\mathbf{y}_t - \mu_t \mid \mathbf{Y}^{(1)}, \mathbf{Z}, \mathbf{W}^1, ..., \mathbf{W}^c\right] \\
&= \mathbf{D}_t \left[\begin{array}{c} \mathbf{y}_t^{(1)} - \mu_t^{(1)} \\ \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\left(\mathbf{y}_t^{(1)} - \mu_t^{(1)}\right) \end{array} \right] \\
\mathbf{\Lambda}_t &= Var_{\Psi^{(k)}}\left[\mathbf{y}_t - \mu_t \mid \mathbf{Y}^{(1)}, \mathbf{Z}, \mathbf{W}^1, ..., \mathbf{W}^c\right] \\
&= \mathbf{D}_t \left[\begin{array}{cc} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{22} - \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12} \end{array} \right] \mathbf{D}_t' \\
\mathbf{\Omega}_t^{(11)} &= E_{\Psi^{(k)}}\left[\mathbf{e}_t^{(1)} \mid \mathbf{Y}^{(1)}\right] \cdot E_{\Psi^{(k)}}\left[\mathbf{e}_t^{(1)} \mid \mathbf{Y}^{(1)}\right]' + Var_{\Psi^{(k)}}\left[\mathbf{e}_t^{(1)} \mid \mathbf{Y}^{(1)}\right] \\
\mathbf{w}_t^{j,T} &= E_{\Psi^{(k)}}\left(\mathbf{w}_t^j \mid \mathbf{Y}^{(1)}\right); 1 \leq j \leq c \\
\mathbf{A}_t^{j,T} &= Var_{\Psi^{(k)}}\left(\mathbf{w}_t^j \mid \mathbf{Y}^{(1)}\right); 1 \leq j \leq c
\end{aligned}
$$

Moreover, $\mathbf{z}_t^T = E_{\Psi^{(k)}}\left(\mathbf{z}_t \mid \mathbf{Y}^{(1)}\right)$ and $\mathbf{P}_t^T = Var_{\Psi^{(k)}}\left(\mathbf{z}_t \mid \mathbf{Y}^{(1)}\right)$ are given by the

Kalman smoother output and

$$\mathbf{S}_{11} = \sum_{t=1}^{T} \mathbf{z}_t^T \left(\mathbf{z}_t^T\right)' + \mathbf{P}_t^T;\ \mathbf{S}_{10} = \sum_{t=1}^{T} \mathbf{z}_t^T \left(\mathbf{z}_{t-1}^T\right)' + \mathbf{P}_{t,t-1}^T;\ \mathbf{S}_{00} = \sum_{t=1}^{T} \mathbf{z}_{t-1}^T \left(\mathbf{z}_{t-1}^T\right)' + \mathbf{P}_{t-1}^T$$

The maximization step of the EM algorithm involves the minimization

$$\Psi^{(k+1)} = \arg\min_{\Psi} Q\left(\Psi, \Psi^{(k)}\right)$$

The estimates $\hat{\theta}^{(k+1)} = \left\{\hat{\theta}^1, ..., \hat{\theta}^c\right\}^{(k+1)}$ and $\hat{\mathbf{V}} = \left\{\hat{\mathbf{V}}^1, ..., \hat{\mathbf{V}}^c\right\}^{(k+1)}$ are obtained by numerical minimization. The close form solutions for $\hat{\mathbf{G}}^{(k+1)}$ and $\hat{\mathbf{\Sigma}}_{\eta}^{(k+1)}$ are already given in Fassò and Finazzi (2011) while the solution for the remaining parameters are obtained by solving $\frac{\partial Q\left(\Psi, \Psi^{(k)}\right)}{\partial \Psi} = 0$ and they are

$$\begin{aligned}
\left(\hat{\sigma}_{i,\varepsilon}^2\right)^{(k+1)} &= \frac{tr\left(\sum_{t=1}^{T} \mathbf{\Omega}_t|_{i,i}\right)}{Tn_i} \\
\hat{\beta}^{(k+1)} &= \left[\sum_{t=1}^{T} \left(\mathbf{U}_{\mathbf{x},t}' \mathbf{U}_{\mathbf{x},t}\right)\right]^{-1} \cdot \left[\sum_{t=1}^{T} \mathbf{X}_{\mathbf{x},t}' \left(\mathbf{e}_t^T + \mathbf{U}_{\mathbf{x},t}' \beta^{(k)}\right)\right] \\
\hat{\gamma}_i^{(k+1)} &= \frac{tr\left[\sum_{t=1}^{T} \left(\mathbf{F}_t^T - \mathbf{G}_t^T - \mathbf{H}_t^T\right)\right]}{tr\left[\sum_{t=1}^{T} \mathbf{U}_{\mathbf{w},t} \left(\mathbf{w}_t^{i,T} \cdot \left(\mathbf{w}_t^{1,T}\right)' + \mathbf{A}_t^{i,T}\right) \mathbf{U}_{\mathbf{w},t}'\right]}
\end{aligned}$$

for each $1 \leq i \leq q$, with $\mathbf{\Omega}_t|_{i,i}$ the $i-th$ diagonal block of $\mathbf{\Omega}_t$. Moreover

$$\begin{aligned}
\mathbf{F}_t^T &= \left(\mathbf{e}_t^T + \gamma_i \mathbf{U}_{\mathbf{w},t}^i \mathbf{w}_t^{i,T}\right) \left(\mathbf{w}_t^{i,T}\right)' \left(\mathbf{U}_{\mathbf{w},t}^i\right)' \\
\mathbf{G}_t^T &= 2 \sum_{j \neq i}^{c} \gamma_j \mathbf{U}_{\mathbf{w},t}^i Cov_{\Psi^{(k)}}\left(\mathbf{w}_t^i, \mathbf{w}_t^j \mid \mathbf{Y}^{(1)}\right) \left(\mathbf{U}_{\mathbf{w},t}^j\right)' \qquad (2) \\
\mathbf{H}_t^T &= 2 \mathbf{U}_{\mathbf{z},t} Cov_{\Psi^{(k)}}\left(\mathbf{z}_t, \mathbf{w}_t^i \mid \mathbf{Y}^{(1)}\right) \left(\mathbf{U}_{\mathbf{w},t}^i\right)' \qquad (3)
\end{aligned}$$

and the conditional covariances in (2) and (3) are computed straightforwardly from the multivariate Gaussian distribution of the joint $(\mathbf{y}_t, \mathbf{w}_t, \mathbf{z}_t)$.

References

Fassò A. and Finazzi F. (2011) Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. Environmetrics. In printing.

Likelihood Inference in Multivariate Model-Based Geostatistics [1]

Clarissa Ferrari, Marco Minozzo
University of Verona, marco.minozzo@univr.it

Abstract: Multivariate model-based geostatistics refers to the extension of classical multivariate geostatistical techniques, in particular the classical linear model of coregionalization, to the case of non-Gaussian data. Extensions of this kind are still limited in the statistical literature, mainly for the inferential problems they pose, and almost invariably inference is carried out in a Bayesian context. In this work we present some new results on likelihood inference for the unknown parameters of a hierarchical geostatistical factor model. In particular, we show the implementation of some Monte Carlo EM algorithms and discuss their performances, in particular their sampling distributions, mainly through some simulation studies.

Keywords: Cokriging, Generalized linear mixed model, Markov chain Monte Carlo, Monte Carlo EM, Multivariate geostatistics.

1 Introduction

The classical linear model of coregionalization, or its simpler counterpart, the proportional covariance model, otherwise known as intrinsic correlation model, and the related 'factorial kriging analysis' have become standard tools in many areas of application for the analysis of multivariate spatial data. However, in presence of non-Gaussian data, in particular count or skew data, the use of these geostatistical instruments can lead to misleading predictions and to erroneous conclusions about the underling factors. To cope with these situations, following the proposal put forward in the univariate case by Diggle et al. (1998), and somehow extending the works of Zhang (2007) and of Zhu et al. (2005), we propose in Section 2 a hierarchical multivariate spatial model, built upon a generalization of the classical geostatistical proportional covariance model. Adopting a non-Bayesian inferential framework, and assuming that the number of underlying common factors and their spatial autocorrelation structure are known, in Section 3 we show how to carry out likelihood inference on the parameters of the model by exploiting the capabilities of Markov chain Monte Carlo (MCMC) and Monte Carlo EM (MCEM) algorithms.

[1] We gratefully acknowledge funding from the Italian Ministry of Education, University and Research (MIUR) through PRIN 2008 project 2008MRFM2H.

2 Multivariate Model-Based Geostatistics

Let us consider the following hierarchical extension of the classical geostatistical linear model of coregionalization. Let $y_i(\mathbf{x}_k)$, $i = 1, \ldots, m$, $k = 1, \ldots, K$, be a set of geo-referenced data measurements relative to m regionalized variables, gathered at K spatial locations $\mathbf{x}_k$. These m regionalized variables are seen as a partial realization of a set of m random functions $Y_i(\mathbf{x})$, $i = 1, \ldots, m$, $\mathbf{x} \in \mathbb{R}^2$. For these functions we assume, for any $\mathbf{x}$, and for $i \neq j$,

$$Y_i(\mathbf{x}) \perp\!\!\!\perp Y_j(\mathbf{x}) | Z_i(\mathbf{x}) \quad \text{and} \quad Y_i(\mathbf{x}) \perp\!\!\!\perp Z_j(\mathbf{x}) | Z_i(\mathbf{x}), \tag{1}$$

and, for $\mathbf{x}' \neq \mathbf{x}''$, and $i, j = 1, \ldots, m$,

$$Y_i(\mathbf{x}') \perp\!\!\!\perp Y_j(\mathbf{x}'') | Z_i(\mathbf{x}') \quad \text{and} \quad Y_i(\mathbf{x}') \perp\!\!\!\perp Z_j(\mathbf{x}'') | Z_i(\mathbf{x}'), \tag{2}$$

where $Z_i(\mathbf{x})$, $i = 1, \ldots, m$, $\mathbf{x} \in \mathbb{R}^2$, are mean zero joint stationary Gaussian processes.

Moreover, for any given i and $\mathbf{x}$, we assume that, conditionally on $Z_i(\mathbf{x})$, the random variables $Y_i(\mathbf{x})$ have conditional distributions $f_i(y; M_i(\mathbf{x}))$, that is, $Y_i(\mathbf{x})|Z_i(\mathbf{x}) \sim f_i(y; M_i(\mathbf{x}))$, specified by the conditional expectations $M_i(\mathbf{x}) = \mathrm{E}[Y_i(\mathbf{x})|Z_i(\mathbf{x})]$, and that $h_i(M_i(\mathbf{x})) = \beta_i + Z_i(\mathbf{x})$, for some parameters β_i and some known link functions $h_i(\cdot)$. For instance, we might assume that for some or all i, and for any given $\mathbf{x}$, the data are conditionally Poisson distributed, that is, that

$$f_i(y; M_i(\mathbf{x})) = \exp\{-M_i(\mathbf{x})\}(M_i(\mathbf{x}))^y / y!, \qquad y = 0, 1, 2, \ldots, \tag{3}$$

and that the linear predictor $\beta_i + Z_i(\mathbf{x})$ is related to the conditional mean $M_i(\mathbf{x})$ through a logarithmic link function so that $\ln(M_i(\mathbf{x})) = \beta_i + Z_i(\mathbf{x})$. On the other hand, for the rest of the i, we might assume that, for any given $\mathbf{x}$, conditionally on $Z_i(\mathbf{x})$, the random variables $Y_i(\mathbf{x})$ are Gamma distributed with conditional expectations $M_i(\mathbf{x}) = \mathrm{E}\big[Y_i(\mathbf{x})\big|Z_i(\mathbf{x})\big] = \exp\big\{\beta_i + Z_i(\mathbf{x})\big\} = \nu b$, (here again $h_i(\cdot) = \ln(\cdot)$) and conditional variances $\mathrm{Var}\big[Y_i(\mathbf{x})\big|Z_i(\mathbf{x})\big] = \nu b^2 = \nu^{-1}\exp\big\{2\beta_i + 2Z_i(\mathbf{x})\big\} = \nu^{-1}(M_i(\mathbf{x}))^2$, where $\nu > 0$ and $b > 0$ are parameters, that is, we might assume

$$f_i(y; M_i(\mathbf{x})) = (y^{\nu-1}/\Gamma(\nu))\exp\{-y\nu/M_i(\mathbf{x})\}(\nu/M_i(\mathbf{x}))^\nu, \qquad y > 0. \tag{4}$$

Here the 'shape' parameter ν is constant for $\mathbf{x} \in \mathbb{R}$, whereas the 'scale' parameter b varies over $\mathbb{R}$ depending on the conditional expectation $M_i(\mathbf{x})$. In addition to the Poisson or Gamma distributions, other discrete or continuous distributions could be considered to account for particular set of data.

For the latent part of the model, we adopt the following structure. For the m joint stationary Gaussian processes $Z_i(\mathbf{x})$, let us assume the linear factor model

$$Z_i(\mathbf{x}) = \sum_{p=1}^{P} a_{ip} F_p(\mathbf{x}) + \xi_i(\mathbf{x}), \tag{5}$$

where a_{ip} are $m \times P$ coefficients, $F_p(\mathbf{x})$, $p = 1, \ldots, P$, are $P \leq m$ non-observable spatial components (common factors) responsible for the cross correlation between the variables $Z_i(\mathbf{x})$, and $\xi_i(\mathbf{x})$ are non-observable spatial components (unique factors) responsible for the residual autocorrelation in the $Z_i(\mathbf{x})$ unexplained by the common factors. We assume that $F_p(\mathbf{x})$ and $\xi_i(\mathbf{x})$ are mean zero stationary Gaussian processes with covariance functions $\mathrm{Cov}\big[F_p(\mathbf{x}), F_p(\mathbf{x}+\mathbf{h})\big] = \rho(\mathbf{h})$, and $\mathrm{Cov}\big[\xi_i(\mathbf{x}), \xi_i(\mathbf{x}+\mathbf{h})\big] = \psi_i \rho(\mathbf{h})$, where $\mathbf{h} \in \mathbb{R}^2$, $\rho(\mathbf{h})$ is a real spatial autocorrelation function common to all factors such that $\rho(0) = 1$ and $\rho(\mathbf{h}) \to 0$, as $||\mathbf{h}|| \to \infty$, and ψ_i are non-negative real parameters. We also assume that the processes $F_p(\mathbf{x})$ and $\xi_i(\mathbf{x})$ have all cross-covariances identically equal to zero.

Assuming that the number P of latent common factors and that the spatial autocorrelation function $\rho(\mathbf{h})$ have already been chosen, the model depends on the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\psi})$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^T$, $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_m)^T$, with $\mathbf{a}_i = (a_{i1}, \ldots, a_{iP})$, for $i = 1, \ldots, m$, and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_m)^T$. Let us note that, as the classical linear factor model, our model is not identifiable. However, the only indeterminacy stays in a rotation of the matrix $\mathbf{A}$.

3 Likelihood inference via MCEM

Adopting a non-Bayesian inferential framework, likelihood inference on the parameters of the model would require the maximization, with respect to $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\psi})$, of the likelihood based on the marginal density function of the observations $y_i(\mathbf{x}_k)$. However, since this marginal density is not available, and since the integration required in the E-step of the EM algorithm would not be easy, here, to maximize the log-likelihood, we will resort to the MCEM algorithm (see Wei and Tanner 1990).

Our implementation of the algorithm proceeds as follows. Let us define $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m)$ where $\boldsymbol{\xi}_i = (\xi_i(\mathbf{x}_1), \ldots, \xi_i(\mathbf{x}_K))^T$, $i = 1, \ldots, m$, and $\mathbf{F} = (\mathbf{F}_1, \ldots, \mathbf{F}_P)$ where $\mathbf{F}_p = (F_p(\mathbf{x}_1), \ldots, F_p(\mathbf{x}_K))^T$, $p = 1, \ldots, P$, and let $f(\mathbf{y}, \boldsymbol{\xi}, \mathbf{F}; \boldsymbol{\theta})$ be the joint distribution of the model, that is, the complete log-likelihood, accounting also for the unobserved factors. Assuming that the current guess for the parameters after the $(s-1)$th iteration is given by $\boldsymbol{\theta}_{s-1}$, and that R_s is a fixed positive integer, the sth iteration of the MCEM algorithm involves the following three steps (stochastic, expectation, maximization):

S step – draw R_s samples $(\boldsymbol{\xi}^{(r)}, \mathbf{F}^{(r)})$, $r = 1, \ldots, R_s$, from the (filtered) conditional distribution $f(\boldsymbol{\xi}, \mathbf{F} | \mathbf{y}; \boldsymbol{\theta}_{s-1})$;

E step – compute $Q_s(\boldsymbol{\theta}, \boldsymbol{\theta}_{s-1}) = (1/R_s) \sum_{r=1}^{R_s} \ln f(\mathbf{y}, \boldsymbol{\xi}^{(r)}, \mathbf{F}^{(r)}; \boldsymbol{\theta})$;

M step – take as the new guess $\boldsymbol{\theta}_s$ the value of $\boldsymbol{\theta}$ which maximizes $Q_s(\boldsymbol{\theta}, \boldsymbol{\theta}_{s-1})$.

With R_s very large this procedure approximates the EM algorithm, whereas a simulated annealing version could be obtained by choosing an increasing sequence $R_s \to \infty$, as $s \to \infty$, (see, for instance, Fort and Moulines 2003). The S-step of the

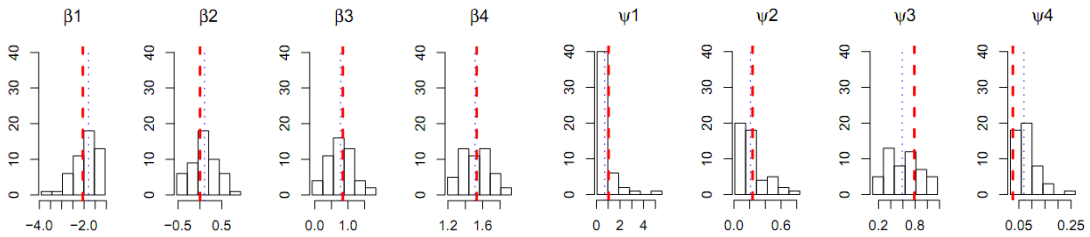


Figure 1: Histograms of the simulated marginal distributions of the MCEM estimator for the 8 parameters of a model with $m = 4$ and one common factor, obtained by running the algorithm over 50 simulated data sets. Dashed lines are the true parameter values; dotted lines are the empirical arithmetic means of the distributions.

algorithm can be dealt with through importance sampling or MCMC techniques, whereas the M-step typically requires the use of numerical routines.

When the matrix $\mathbf{A}$ is known, the complete log-likelihood belongs to the curved exponential family and by choosing an appropriate increasing sequence R_s the algorithm converges to the maximum likelihood estimate (Fort and Moulines 2003). On the other hand, when the matrix $\mathbf{A}$ is unknown, the complete likelihood does not belong any more to the curved exponential family and theoretical convergence properties are not available. However, we show, either in the case in which $\mathbf{A}$ is known or unknown, through some extensive simulation studies, that the MCEM algorithm provides estimates with quite reasonable sampling distributions. For instance, Figure 1 shows the simulated distributions of the MCEM estimator in the case in which $P = 1$ and $\mathbf{A}$ is known.

References

Diggle P. J., Moyeed R. A., Tawn J. A. (1998) Model-based geostatistics (with discussion), *Applied Statistics*, 47, 299–350.

Fort G., Moulines E. (2003) Convergence of the Monte Carlo expectation maximization for curved exponential families, *The Annals of Statistics*, 31, 1220–1259.

Wei G. C. G., Tanner M. A. (1990) A Monte Carlo implementation of the EM algorithm and poor man's data augmentation algorithm. *Journal of the American Statistical Association*, 85. 699–704.

Zhang H. (2007) Maximum-likelihood estimation for multivariate spatial linear coregionalization models, *Environmetrics*, 18, 125–139.

Zhu J., Eickhoff J. C., Yan P. (2005) Generalized linear latent variable models for repeated measures of spatially correlated multivariate data, *Biometrics*, 61, 674–683.

A system for on-line measurement of key soil properties

Abdul Mounem Mouazen, Boyan Kuang & Mohammed Zakaullah Quraishi
Environmental Technology and Science Department, Cranfield University, Bedfordshire, MK43 0AL, United Kingdom, e-mail: a.mouazen@cranfield.ac.uk

Abstract: This paper reports on on-line measurement of total nitrogen (TN), organic carbon (OC), moisture content (MC) and bulk density (BD) which was carried out in three European farms. The measurement system consists of a multiple sensor platform, which includes a mobile, fibre-type, visible and near infrared (vis-NIR) spectrophotometer, a draught (D) and a depth (d) sensor. The prediction models developed were tested in three fields in Denmark, Czech Republic and UK. Results revealed that the measurement accuracy was very good for TN (RPD = 2.33 - 2.38), very good/excellent for OC (RPD = 2.31 - 2.52) and excellent for MC (RPD = 3.16 - 3.25). A reasonably good correlation between the measured and on-line predicted BD (R^2 = 0.56 – 0.73) was obtained. The on-line measured maps were similar to those developed with traditional laboratory method of soil analysis.

Keywords: On-line measurement, sensor fusion platform, soil properties.

1. Introduction

Proximal soil sensing becomes one of the main requirements for successful implementation of precision agriculture. Among others, vis-NIR on-line sensors are characterised to be fast, robust, cost effective and environment friendly soil spatial variability detecting techniques. Among few vis-NIR on-line soil sensors available today, the system of Mouazen (2006) is a sensor fusion platform that enables measurement of several soil properties with the vis-NIR spectroscopy in addition to data fusion algorithm for the measurement of soil BD (Mouazen and Ramon, 2006). This paper aims at reporting on automatic data collection of multiple soil properties at farm scale using a sensor fusion platform (Mouazen, 2006) in three fields across three European farms.

2. Material and methods

The on-line measurement system designed and developed by Mouazen (2006) was used. An AgroSpec mobile, fibre type, vis-NIR spectrophotometer (Tec5 Technology for Spectroscopy, Germany) with a measurement range of 305-2200 nm was used to measure soil spectra in reflectance mode. The spectrometer was IP 64 protected for harsh working environments. A shear beam load cell (9 tonne capacity) for the measurement of D and draw wire linear sensor (connected to a depth wheel) for the measurement of subsoiler d were used. A single A DGPS (EZ-Guide 250, Trimble, USA) was used to record the position of the on-line measurements with sub-meter accuracy. A Panasonic semi-rugged laptop was used for data logging and communication. The spectrometer system, laptop and

DGPS were powered by the tractor battery. A total of 3 fields were measured in three farms in Czech Republic, Denmark and the UK in summer 2010. In each field, blocks of 150 m by 150 m, covering about 2 ha of land were measured. About 2 or 3 soil samples were collected from each measurement line with 28 to 48 soil samples collected from each field.

Soil chemical analyses and optical measurements were carried at Cranfield University. Soil OC and TN were measured by a TrusSpecCNS spectrometer (LECO Corporation, St. Joseph, MI, USA) using the Dumas combustion method. Soil MC was determined by oven drying of soil at 105 ºC for 24 h. Having MC measured, BD was calculated for all samples.

Each soil sample was dumped into a glass container and mixed well. Big stones and plant residue were excluded. Then each soil sample was placed into three Petri dishes, which were 2 cm in depth and 2 cm in diameter. The soil in the Petri dish was shaken and pressed gently before levelling with a spatula. The soil samples were scanned with the same AgroSpec spectrophotometer used for on-line measurement. A 100 % white reference was used before scanning. A total of 10 scans were collected from each cup, and these were averaged in one spectrum.

Soil spectra were first reduced to 371 - 2150 nm to eliminate the noise at both edges of each spectrum. Spectra were further reduced by averaging three successive points in the vis range, and 10 points in the NIR range. The Savitzky-Golay smoothing, maximum normalisation and first derivation were successively implemented using Unscrambler 7.8 software (Camo Inc.; Oslo, Norway). The pre-treated spectra and the laboratory chemical measurement values were used to develop calibration models for OC, TN and MC.

General calibration models developed previously for TN, OC and MC, using 480 soil samples collected from 4 farms across 4 European countries were used in this study. Out of 188 samples collected from the three fields in Europe, 63 samples were randomly selected for the calibration and the remaining 125 samples were used as independent validate set. However, the range of variation of each property for both the calibration and validation sets was almost identical. The calibration samples were spiked into the original calibration set of the general calibration models developed for European soils. The calibration spectra were subjected to a partial least squares regression (PLSR) with the leave-one-out cross validation using the Unscrambler 7.8 software.

A model (Eqn. 1) to predict BD based on measured D, d and MC was developed by Mouazen et al. (2009), which is valid for sand, loam, silt loam and silt loam/silt soils. Equation (1) was used to predict BD in this study.

$$BD = \left(\sqrt[3]{\frac{D + 21.36\,MC - 73.9313\,d^2}{1.6734}}\right) \times (1.240 - 0.592MC - 0.000792clay) \quad (1)$$

Where D is subsoiler draught [kN], MC is gravimetric moisture content [kg kg^{-1}], d is cutting depth [m] and BD is bulk density [Mg m^{-3}] and clay is expressed in %.

3. Results and discussion

The measurement campaign proved that the sensor fusion platform can provide simultaneous measurement of several soil properties. This platform enabled the collection of around 3000 data points from each field with an average of around 2 points per meter travel distance. The chemical analysis values of the manually collected samples were compared with the on-line predicted concentration values using vis-NIR spectra collected in the same positions. Table 1 summarises the results of general model accuracy and on-line validation results. The general calibration models refer to those developed originally using 480 samples with the spiked 65 samples from the three fields measured in this study. The validation of the on-line measured data was based on 125 samples collected from the three fields for validation. Examining the ratio of prediction deviation (RPD), which is the standard deviation (SD) divided by root mean square error of prediction (RMSEP), revealed that RPD values were above 2 for all soil properties in all fields. An RPD value between 1.5 and 2 and between 2.0 and 2.5 indicates good and very good quantitative model predictions, respectively. Values above 2.5 indicate excellent prediction results (Viscarra Rossel et al., 2006). Adopting this classification system of the prediction accuracy reveals that prediction performance for TN, OC and MC is very good to excellent performance (Table 1).

		OC		**TN**		**MC**		**BD**
Validation		RMSEP %	RPD	RMSEP %	RPD	RMSEP %	RPD	R^2
	CZ field	0.07	2.33	0.007	2.52	0.72	3.16	0.56
	DK field	0.05	2.38	0.004	2.47	0.37	3.25	0.72
	UK field	0.09	2.38	0.008	2.31	0.59	3.25	0.73
Calibration models		0.104	2.89	0.009	2.93	1.05	4.32	-

Table 1. Calibration and field validation results of the on-line measurement in 3 fields

The prediction of BD with the on-line sensor fusion platform provided reasonably good accuracy (Table 1). Although the model was developed for fields in Belgium (Mouazen et al., 2009), the R^2 values shown in Table 1 confirm that the model (Eqn 1) can be expanded to other fields across the European countries considered in this study, as long as the same soil textures to those used to build the original calibration model (Eqn. 1) are used.

Using an ArcGIS 10 (ESRI, USA) mapping software, maps for the selected soil properties were developed. The inverse distance weighting (IDW) method was used for the spatial interpolation. In order to allow for useful comparisons between reference and on-line measured maps, the same number of classes (seven classes) was considered for all maps. Figure 1 compares between maps of on-line and laboratory measured TN in the UK field, taken as an example. The same number of samples (18 validation samples) was used to develop both maps. A comparison between maps of measured and predicted TN shows large similarity, which was also achieved for OC, MC and BD (data is not shown).

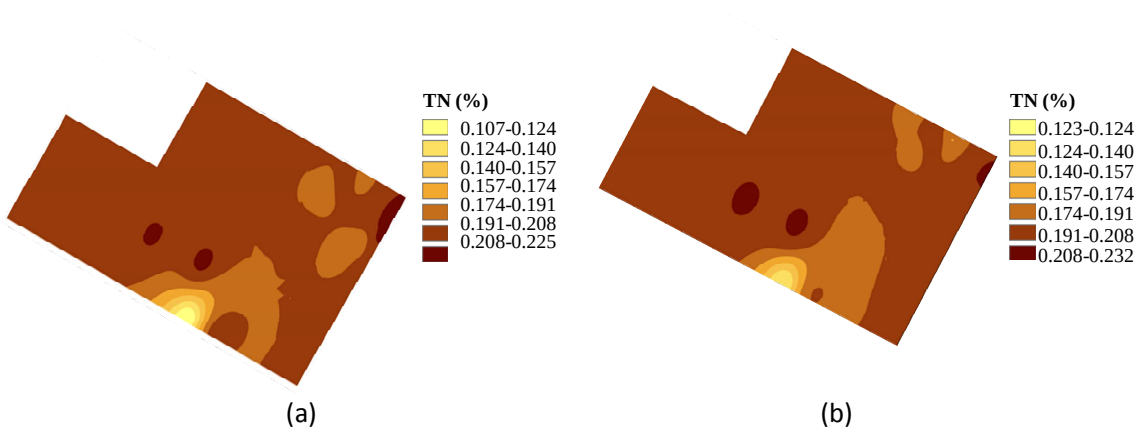


Figure 1. Comparison between laboratory (a) and on-line (b) measured total nitrogen (TN) in the UK field, using 18 validation samples

4. Conclusions

This paper reports on the performance of on-line sensor fusion platform for the measurement of multiple soil properties at farm scale in Europe. Results reported in this study allow the following conclusions to be drawn:

1- The on-line sensor fusion platform enabled the collection of large data points (about 3000 points per field).

3- The accuracy of on-line measured OC, TN and MC was classified as very good to excellent prediction performance. Reasonable good measurement of BD was reported.

References:

Mouazen, A. M., De Baerdemaeker, J., Ramon, H. (2005). Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer. *Soil & Tillage Research*, 80, 171-183.

Mouazen, A. M. (2006). Soil Survey Device. International publication published under the patent cooperation treaty (PCT). World Intellectual Property Organization, International Bureau. International Publication Number: WO2006/015463; PCT/BE2005/000129; IPC: G01N21/00; G01N21/00.

Mouazen, A. M., Ramon, H. (2006). Development of on-line measurement system of bulk density based on on-line measured draught, depth and soil moisture content. *Soil & Tillage Research*, 86, 218-229.

Mouazen, A. M.; Ramon, H., (2009). Expanding implementation of an on-line measurement system of topsoil compaction in loamy sand, loam, silt loam and silt soils. *Soil & Tillage Research*, 103, 98-104.

Viscarra Rossel, R. V., McGlyn R. N., McBratney, A. B. (2006). Determining the composition of mineral-organic mixes using UV–vis–NIR diffuse reflectance spectroscopy. *Geoderma*, 146, 403-411.

Modified Hot-Spot analysis for spatio-temporal analysis: a case study of the leaf-roll virus expansion in vineyards [1]

Yafit Cohen 1
Agricultural Research Organization, Israel, Yafitush@volcani.agri.gov.il

Rakefet SHARON, Tamar SOKOLSKY*, Tirza ZAHAVI 2
Northern R&D, Israel
*The Robert H. Smith Faculty of Agriculture, Food and environment

Abstract: Given a set of geo-referenced data points, the Getis–Ord Gi* statistic identifies hot-spots of points with values higher in magnitude than one might expect by a random chance. This tool works by looking at each data feature and its neighboring features in comparison to the overall spatial distribution of the phenomenon explored. If the difference between the local sum for a feature and its neighbors is highly larger than expected (the overall sum) a hot-spot is accepted. Leaf-roll virus (LRV) in vineyards appears in clusters which expand from year to year when no pest control is carried out. Exploring the spatio-temporal expansion of hot-spots of the LRV is limited with the Gi* statistics since relative hot-spots are accepted according to the infestation level of a specific year. A modified Gi* was developed which identifies year-to-year hot-spots which are relative to a year of reference. LRV symptoms were mapped yearly in a vineyard from 2005 to 2010. The Gi* indicated for a northern hot-spot only in 2007 and a southern one in 2009. Using the modified Gi* with 2005 as a year of reference, the northern and the southern clusters were identified in 2006 and 2007, respectively.

Keywords: Getis–Ord Gi*, Leaf-roll virus, Spatio-temporal dynamics

1. Introduction

Leafroll is one of the most widespread viral diseases of grapevine. Leafroll disease is an economically important graft-transmissible disease of grapevines and occurs in all grapevine-growing countries. Although grapevine leafroll virus (LRV) can affect the growth, development, longevity and yield of the vines, its most serious effect is on lowering the sugar content and raising the acidity of must. In the field, the spread of LRV associated with particular insect vectors has been reported in several countries, including Israel (Cabaleiro and Segura, 2006; Tanne et al., 1989). In vineyards with available virus inoculum and mealybugs present, LRV spreads quite quickly from vine to vine (Cabaleiro et al., 2008). Study of the spatio-temporal dynamics of the leafroll-infected vines in the vineyard scale may be helpful to determine whether or not spread

[1] Chief Scientist of the Ministry of Agriculture, Israel, research no 132-1502-09

.

of the viruses is occurring, and the best control measures to take. The spreading via the mealybugs creates clusters of infested vines which expanding from year to year. Local spatial statistics may assist with the identification of the infestation clusters. Local spatial statistics identifies those clusters with values higher in magnitude than is expected to be found by random chance. The Getis–Ord Gi* hotspot cluster statistic is one of the many possible approaches used for local spatial analysis (Getis and Ord, 1996). The Gi* statistic measures the degree of spatial clustering of a local sample and how different it is from the expected value which is the mean of the whole data set. Study of the annual expansion of the LRV can use the annual maps of hot spots and follow their expansion. Yet, since the Gi* statistic is a relative measure to the overall infestation level mean in a particular year and since the mean infestation increases annually, the discovery of hot spots is limited. The objective of this paper is to describe a modification of the Gi* statistic which enables a spatio-temporal analysis of the LRV hot-spots expansion.

2. Materials and Methods

The Gi* statistic

The Gi* statistic measures the degree of spatial clustering of a local sample and how different it is from the expected value (Equation 1). It is calculated as the sum of the differences between values in the local sample and the mean, and is standardized as a z-score with a mean of zero and a standard deviation of 1:

$$G_i^*(d) = \frac{\sum_j w_{ij}(d)x_j - W_i^*\bar{x}^*}{s^*\sqrt{\frac{(nS_{1i}^*) - W_i^{*2}}{n-1}}}$$

Equation 1:

where i is the centre of the local neighborhood, d is the lag distance (radius), w_{ij} is the weight for neighbor j from location i, n is the number of samples in the data set, W_i* is the sum of the weights, S_{1i}* is the number of samples within d of the central location, x* is the mean of the whole data set, and s* is the standard deviation of the whole data set. The Gi* statistic is two-tailed, so a score of ±2 represent strong clustering, as 95% of the data under a normal distribution should be within 2 standard deviations of the mean. Values between ±2 may be interpreted as weakly clustered, with values being less than 2 standard deviations away from what one would expect if there were no spatial clustering (Laffan, 2006). While positive values of Gi* represent clusters that are, on average, greater than the mean (Hot-spots) the negative values represent clusters that are less than the mean (Cold-spots). The Gi* statistic is a relative measure and the existence of hot spots of LRV infested vines is highly depended on the overall mean. Since LRV is expanding annually the overall mean increases and local expansions of infestation might not be observed. This attribute limits the use of the Gi* statistic for spatio-temporal analysis of the LRV expansion. A modified Gi* is suggested which calculates Gi* using a pre-defined overall mean (x*) and standard deviation of (s*) infestation which function as a common baseline. In this way local expansion of hot spots may be observed.

To explore the potential in using the modified Gi* (*m*Gi*) real data of LRV infested vines were used. Ten rows in a vineyard in the Golan Heights in Israel were monitored

for LRV symptoms from 2005 to 2010 (overall of 1142 vines). The study area was divided into 4X4 meters cells (3-4 vines). Each cell was set with the number of infested vines inside it in each year. Gi* statistic was calculated for a radius of 24 meters (*d* in equation 1) for each year allowing at least 30 data features which are required for a valid analysis (Getis and Ord, 1996). No weights were set to the data. For the calculation of *m*Gi* the overall mean (*x**) which is the proportion of the infested vines the standard deviation (*s**) of the year 2005 were used. The calculations were made in Matlab, transferred into shapfiles and mapped in ArcGIS 9.3.1.

3. Results

Table 1 shows the average and standard deviation of LRV infestation levels in the years 2005-2010. The infestation level increased over the years. A notable increase was between the years 2008 and 2009 and between 2009 and 2010.

Year	Infestation mean	STD
2005	0.162	0.421
2006	0.219	0.492
2007	0.266	0.592
2008	0.320	0.694
2009	0.563	0.964
2010	0.823	1.159

Table 1: The mean and standard deviation of infestation levels in the years 2005-2010.

Figure 1 presents maps of number of LRV infested vines in 2005, 2007 and 2009. Visual interpretation of the maps indicates for clustered infestation distribution. There was a group of infested vines in the northern part of the vineyard which expanded along the years. In 2009 another distinctive group of infested vines was located in the southern part.

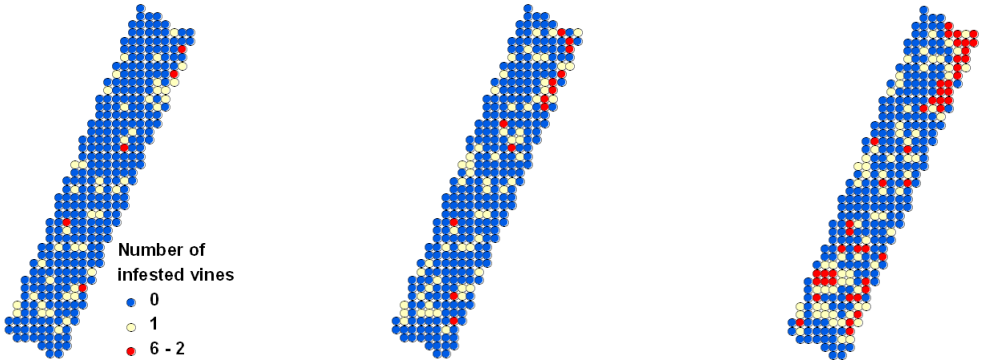


Figure 1: Maps of infested vines in 2005 (left), 2007 (center), and 2009 (right)

Figure 2 presents maps of hot and cold spots in 2005, 2007 and 2009 using the Gi*. The hot spot analysis using the Gi* generally agrees with the visual analysis. Nonetheless, an increase in cold spots was also observed and this is despite the fact that the average infestation level was doubled from 2007 and 2009 (Table 1). Additionally, the southern hot spot in 2009 is relatively small in comparison with the infestation map of 2009. The

respective maps created based on the *m*Gi* calculations (Figure 3) shows the advantages and weakness of the *m*Gi*. In 2007 the *m*Gi* discovered a much larger cluster in the northern part on comparison with the Gi*. The *m*Gi* also discovered the southern cluster already in 2007 while by the Gi* this cluster was discovered only in 2009. On the other hand, in 2009 almost all the vineyard was mapped as hot spot.

4. Concluding remarks

The *m*Gi* may be used for spatio-temporal study of LRV expansion but better reference values for its calculations need to be defined.

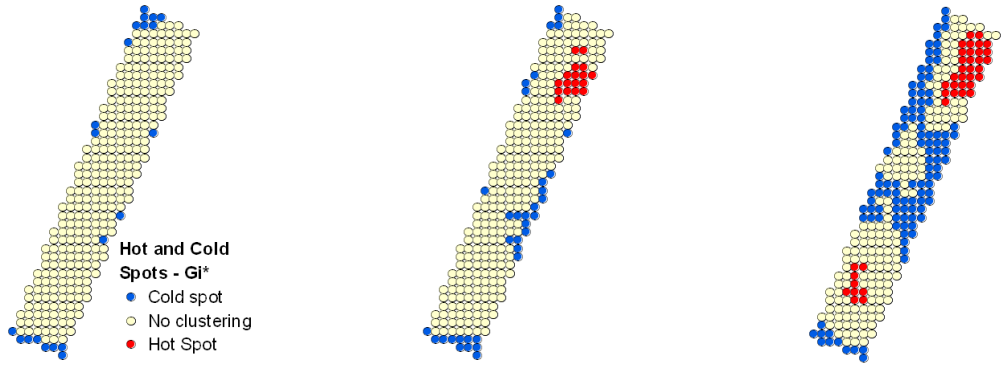


Figure 2: Maps of hot and cold spots using Gi* in 2005 (left), 2007 (center), and 2009 (right)

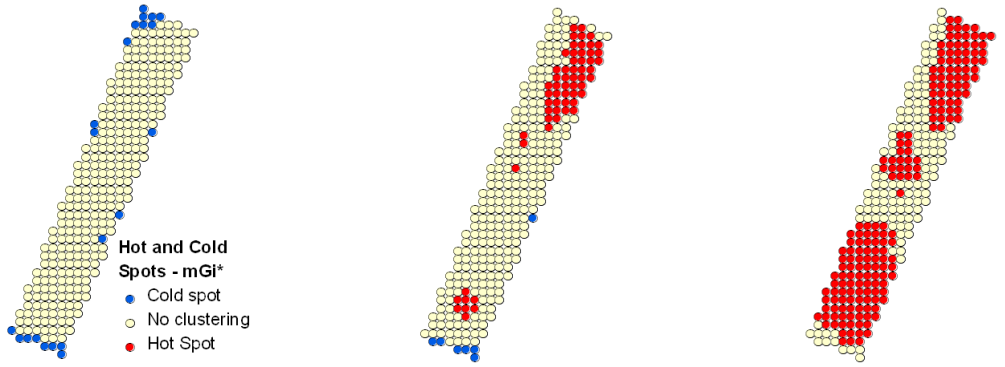


Figure 3: Maps of hot and cold spots using *m*Gi* in 2005 (left), 2007 (center), and 2009 (right)

References

Cabaleiro, C., Couceiro, C., Pereira, S., Cid, M., Barrasa, M., and Segura, A. (2008). Spatial analysis of epidemics of Grapevine leafroll associated virus-3. *European Journal of Plant Pathology* **121**, 121-130.

Cabaleiro, C., and Segura, A. (2006). Temporal Analysis of Grapevine leafroll associated virus 3 Epidemics. *European Journal of Plant Pathology* **114**, 441-446.

Getis, A., and Ord, J. (1996). Local spatial statistics: an overview. *In* "Spatial Analysis: Modeling in a GIS environment " (P. Longley and M. Batty, eds.), pp. 269-294. GeoInformation International, Cambridge.

Laffan, S. W. (2006). Assessing regional scale weed distributions, with an Australian example using Nassella trichotoma. *Weed Research* **46**, 194-206.

Tanne, E., Ben-Dov, Y., and Raccah, B. (1989). Transmission of closterolike particles by mealybugs (Pseudococcidae) in Israel. *In* "Proceedings 9th Meeting ICVG", pp. 71-73, Kiryat Anavim.

Multimodal remote sensing for enhancing detection of spatial variability in agricultural fields[1]

Victor ALCHANATIS, Avi COHEN, Yafit COHEN
Institute of Agricultural Engineering, Agricultural Research Organization, The Volcani Center, POB 6, Bet Dagan 50250, Israel, victor@volcani.agri.gov.il

Ofer Levi
Dept. of Industrial Management, Ben Gurion University, Be'er Sheva, Israel

Amos NAOR
Golan Research Institute, University of Haifa, P.O. Box 97, Kazrin 12900, Israel

Abstract: Detection of variability in agricultural fields depends on the spatial scale of the observed variable. Plant water status can be evaluated using thermal IR images that can provide valuable information on the water status, whereas visible RGB images can provide detailed information on the plants' color, which is not a good indicator of the water status. The informative mode (thermal IR images) has coarse resolution, as opposed to the excessive resolution of the less informative mode (visible RGB). In the present study, we present a method to enhance the information obtained from the thermal IR mode, by combining information from the visible RGB mode. We propose to un-mix the temperature of objects in the thermal images based on the information extracted from the high resolution RGB image.

Keywords: thermal un-mixing, end members, segmentation.

1. Introduction

One limitation in the use of thermal imaging for determining crop temperature and crop water status is that a pure sunlit canopy temperature is needed, and inclusion of shaded leaves and soil background can result in false detection of water stress. To overcome this, high spatial resolution thermal images were combined with images in the visible and NIR ranges (Moran et al., 1994; Clarke, 1997, Möller et al., 2007; Sela et al., 2007). Our group (Sela et al., 2007 and Möller et al., 2007) worked on high resolution images and used the images in the visible range to exclude soil and shaded leaves, resulting in high correlations between the calculated crop water stress index (CWSI) and stomatal conductance. One of the main conclusions from the last works (Sela et al., 2007 and Möller et al. 2007) was that the high spatial resolution (~500 pixels per leaves) enabled proper selection of sunlit leaves. But, the size of the image (6 X 6 m) was impractical for production of crop water status maps on a commercial scale. A much larger image size can be obtained from airborne photography and the increased pixel size introduces new challenge for extracting sunlit leaf temperature by un-mixing a mixture of sunlit and shaded leaves and bare soil that are included in a single pixel.

[1] This research was supported by grant of the Israeli Ministry of Science

Extracting sunlit leaf temperature can be performed by un-mixing. A leading approach for coarse resolution images is the theoretical Vegetation Index-Temperature (VIT) trapezoid (Moran et al. 1994). The VIT trapezoid is the shape obtained when plotting surface composite temperature (Ts) minus air temperature (Ta) as a function of fractional vegetation cover. Theoretically, all variations of crop water stress for different vegetation cover should fall within this trapezoid. Clarke (1997) further showed that the trapezoid could be divided into plant stressed and unstressed regions. The objective of the current work is to develop new un-mixing approach to extract sunlit leaf temperatures using multimodal images (visible RGB and thermal infra red). In the current work we introduce a method that is based on segmentation of high resolution RGB images to a number of end members, and subsequent computation of the end-members' temperature using statistical un mixing methods.

Spectral un-mixing is a common method used to extract pure spectral signatures of objects that are in a mixed environment, and their spatial dimensions are less than the smallest detectable object. In this common case, the additional information that must be provided is the end members, i.e. the characteristics of all the objects that are known to be mixed in that pixel. Based on that additional information, the proportion of each specific material in the pixel can be derived. Our case is different from the common un-mixing problem since we do not know the characteristics of the end members, i.e. we do not know the temperature of the objects that are in the same pixel with the sunlit leaf that we want to measure. Instead, in our case the end-members, their proportion on each pixel and their characteristics will be extracted from multimodal images (visible RGB and thermal infra red).

2. Materials and Methods

High spatial resolution images in the thermal and visible range were acquired around noon time in almond trees under five irrigation treatments in Kibutz Lavee, Israel. A 320x240 pixel microbolometer radiometer (FLIR, SC2000) was used for acquiring thermal infra red images, and a high resolution (8Mpixels) RGB camera (SONY F828) was used for the visible range. Images were acquired from a crane, about 20-25 m above the canopy.

The proposed un-mixing methodology consists of the following steps:

a) The first issue that has to be addressed is the co-registration of the images from the two modes (visible RGB and thermal infra red). In conventional systems, the images are produced by the same sensor and are therefore aligned. The use of two different sensors yields two images of the same scene from slightly different viewing angles, different optical lenses, and different acquiring sensors. Proper alignment of the two images is essential prior to the application of any un-mixing procedure. Alignment or co-registration of images is typically performed with images from the same source. In these cases, correlation based procedures usually yield satisfactory results. But images that carry different basic information cannot be co-registered using conventional techniques. We have developed methodology for multi modal image registration based on mutual information (Wachs et al., 2007). In this work we used mutual information for multi-modal image registration, and manually adjusted the registration were errors were found.

b) from high resolution RGB we obtained the proportions of sunlit and shaded soil and leaves. The RGB images were segmented and classified into 3 end members: sunlit

leaves, sunlit soil and soil in the shadow. Segmentation was performed using the spectral angle mapper (SAM) and the pixels were divided into the 3 end members.

c) for each mixed thermal pixel a linear equation that describe the relationship between the temperatures of the objects and the mixed temperature is defined.

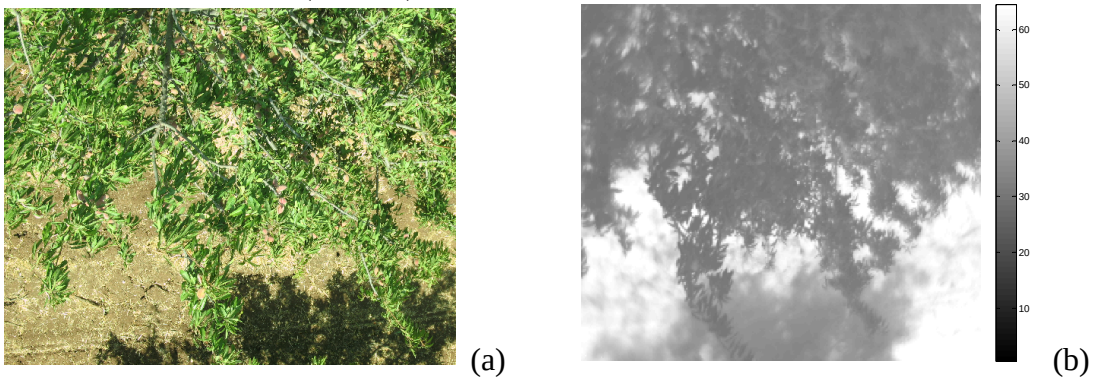
$$T_{i,j} = f_1 T_{ss} + f_2 T_{shs} + f_3 T_l \qquad (1)$$

where: $T_{i,j}$ is the temperature of the mixed pixel (i,j), f_1 and T_{ss} are the proportion and the temperature of sunlit soil respectively, f_2 and T_{shs} are the proportion and the temperature of shaded soil respectively and f_3 and T_l are the proportion and the temperature of the leaves respectively. The solution of the set of linear equations for all the pixels in the image, is the estimated temperature of the end members.

Thirty five images were analyzed, and the canopy temperature was estimated using the proposed un-mixing procedure. Leaf temperatures of pure thermal pixels were manually extracted, and compared to the leaf temperature computed with the proposed un mixing model. Paired t-test analysis was performed in Matlab®. (The Mathworks, US)

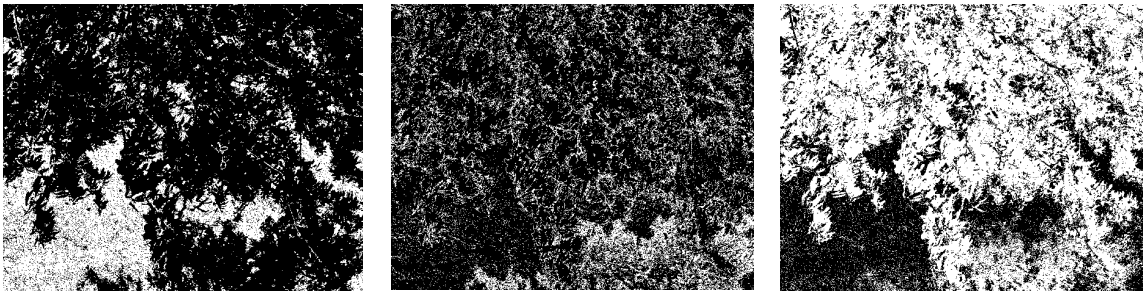
3. Results

Figure 1a shows a sample high resolution RGB image and 1b the corresponding TIR image. Figures 2a,b,c show the segmentation result to three end members, sunlit soil, shaded soil and leaves, respectively. Table 1 shows a comparison of the average leaf temperatures manually extracted for each irrigation treatment and the average leaf temperatures computed with the proposed un-mixing model. The average deviation is approximately 0.4 ^{o}C. Statistical paired t-test for comparison between the manually measured and un-mixing extracted temperatures showed that there is no significant difference between them ($\alpha=0.01$).



(a) (b)

Figure 1: Sample high resolution RGB image (a), and the corresponding TIR image (b).



(a) (b) (c)

Figure 2 Segmentation to three end members: (a) sunlit soil, (b) shades, and (c) leaves.

Irrigation treatment	Manually measured temperature [^{o}C]	Un-mixing extracted temperature [^{o}C]	Number of samples
High stress	33.7 (0.7)	33.3 (0.6)	9
Moderate stress	32.1 (1.0)	31.9 (0.9)	6
Low stress	32.7 (1.3)	32.5 (1.1)	8
Farmer's practice	30.4 (1.4)	30.2 (1.3)	4
No stress	28.8 (0.8)	28.9 (0.8)	8

Table 1: Comparison of the average leaf temperatures manually extracted for each irrigation treatment and the average leaf temperatures computed with the proposed un-mixing model. Numbers in brackets depicts the standard deviation of the sample.

4. Concluding remarks

The proposed algorithm successfully segmented the high resolution images into three end members and subsequently extracted their temperatures. This method can be used to produce high resolution water status maps. The information provided by these maps will be much more detailed that what the growers are used to - it will allow growers to adjust irrigation rates at high resolution (precision irrigation) when the irrigation equipment allows it – this means that a higher proportion of the orchard will be irrigated close to optimum, i.e. the highest, water use efficiency.

References

Clarke T. R. (1997). An empirical approach for detecting crop water stress using multispectral airborne sensors. *Hortechnology* 7, 9–16.

Moran M..S., Clarke T..R., Inoue Y., Vidal A. (1994). Estimating crop water deficit using the relation between surface-air temperature and spectral vegetation index. *Remote Sensing of Environment* 49, 246-263.

Möller, M., V. Alchanatis, Y. Cohen, M. Meron, J. Tsipris, A. Naor, V. Ostrovsky, M. Sprintsin, S. Cohen. (2007). Use of thermal and visible imagery for estimating crop water status of irrigated grapevine. *Journal of Experimental Botany*, 58(4):827-838.

Sela, E., Y. Cohen, V. Alchanatis, Y. Saranga, S. Cohen, M. Möller, M. Meron, A. Bosak, J. Tsipris and V. Orlov. (2007). Thermal imaging for estimating and mapping crop water stress in cotton. *European Conference in Precision Agriculture*, June 2007, Skiathos, Greece. Pages: 365-371.

Wachs. J., H., Stern, T., Burks and V. Alchanatis, 2007. Multi-modal Registration Using a Combined Similarity Measure, *12th Online World Conference on Soft Computing in Industrial application*, October 16-26, 2007.

The Use of the Geoadditive Model with Interactions in a Precision Agriculture Context: a Comparison of Different Spatial Correlation Structures

Barbara Cafarelli, Corrado Crocetta
Università degli Studi di Foggia, b.cafarelli@unifg.it

Annamaria Castrignanò
CRA, Bari

Abstract: Accelerated land degradation is mostly human induced and occurs in all eco-regions regardless of social, economic and political conditions. Precision Agriculture is an ecological management strategy based on the use of several sources of information in order to support decisions concerning the agricultural practice. In this context, the use of methodologies, taking into account spatial and temporal variability associated to every aspect of agricultural production processes, can improve crop yields and environmental quality. In this paper, a geoadditive model with interactions is proposed to analyse the nonlinear relations between an indicator of durum wheat production with other crop features with the aim of considering explicitly the spatial dependence and the temporal variation in production.

Keywords: Geoadditive Model, Matérn family, Spatial Correlation Structures, Cross-validation, Additive model with Interactions

1. Introduction

Precision Agriculture or site-specific crop management is a means of managing spatial and temporal variability of different data types: edaphic (i.e. soil related), anthropogenic, topographic, biological and meteorological factors which are deemed to affect crop yield. The target of Precision Agriculture is to increase crop productivity, optimise inputs, increase farmer's profitability and reduce environmental impact, through the application of variable rate inputs on the basis of the actual local requirements of crop rather than an estimation averaged over the whole field. In this context, defining reliable methods for assessing and predicting within-field variations in soil and crop properties is very important. Effects of the soil's physical and chemical properties on crop yield are predictable and can be mapped relatively easily, whereas effects due to climatic conditions, nutrient deficiency, pests and diseases, being time-dependent, are more difficult to predict. The application of proper statistical models, to assess spatial and temporal variation and predict crop response to site-specific environmental conditions, is then crucial in the perspective of Precision Agriculture. In particular, a geoadditive model with interactions is proposed to analyse the spatial distribution of the harvest index (White, Wilson, 2006), a commonly used indicator of commercial wheat production, and its nonlinear relations with other crop features over two years. The model adopted here is a further development of an extension (Cafarelli

and Castrignanò, 2011) of the original geoadditive model of Kammann and Wand (2003), that explicitly considers data stratified according to crop in two different years. Two geoadditive models with interactions, considering the same response variable and covariates, for the same linear and non-linear relationships between response variable and covariates over time, but differing for spatial correlation structures, are fitted. The selection among the fitted models is done by cross-validation (Carroll and Cressie, 1996).

2. Materials and Methods

The trial was carried out on a 12-ha field cropped with durum wheat (Triticum durum DESF), located at the CER-CRA research centre for cereals, Foggia (41° 27' N, 15° 36' E, 90 m above sea level), south-eastern Italy. The soil was a deep, silty-clay Vertisol of alluvial origin, classified as fine, mesic, Typic, Chromoxerert. The climate was characterized by hot and dry summers and rains concentrated mostly in the winter months. The agricultural trial was carried out during two crop seasons: 2005-2006 and 2007-2008. One-hundred georeferenced measurements of the harvest index (*HI*), number of fertile plants (*FP*) and electrolytic weight (*EW*) were taken for each year. The samples with more than one missing value were discarded leaving only ninety-three and ninety-one georeferenced soil samples to be considered for the first and the second wheat season, respectively. that *HI* had different spatial distributions in the two years, which share a marginal bell shaped distribution. This consideration was supported by a graphical check, which led us to adopt a semi-parametric approach, based on a geoadditive model with interactions. A full representation of the geoadditive model with interactions is:

$$HI_i = \beta_0 + \beta_1 EW_i + g(FP_i) + f_{year}(location\_of\_year_i) + \boldsymbol{\beta}_{\mathbf{x}}\mathbf{x}_i + S(\mathbf{x}_i) + \varepsilon_i, \quad (1)$$

where $i = 1,\ldots,184$ represents the spatial-temporal observation, g and f are smooth functions, $\mathbf{x}_i = (X_i, Y_i)$, in UTM WGS84 coordinate system, is the spatial location of the i-th observation and $S(\mathbf{x}) \sim N(0, \sigma^2_{\mathbf{x}} h_0(r, \nu))$, where $\sigma^2_{\mathbf{x}}$ is the sill, r is the range, ν is the smoothing parameter and $h_0(r,\nu)$ is a Matérn family covariance function used to specify the spatial correlation structures. The exponential and the Gaussian covariance structures were used in the fitted models. This occurred by setting $\nu = 3/2$ or $\nu \to \infty$, respectively, in the function $h_0(r, \nu)$ (Minasny and McBratney, 2005). Independently of the specification of $h_0(\cdot)$, the Gaussian spatial process $S(\cdot)$ is independent of the error term $\boldsymbol{\varepsilon}$ and the additive components. In model 1, the term $f_{year}(\cdot)$ corresponds to the number of spatial locations within a particular year and represents the interaction between the year factor and the overall spatial effect. The relatively small sample size permitted the use of the parsimonious low rank parameterization of model 1 (Hastie, 1996). The choice of linear components was done according to approximated Z-values given by `lme`, while the significance of nonlinear effects, identified with the exploratory data analysis, was assessed by restricted likelihood ratio tests (Kammann and Wand, 2003; Greven et al., 2008; Crainiceanu, 2008; Ruppert et al., 2009). Independently of the spatial correlation structure adopted, the number of nodes for

representing the nonlinear *FP* effect was 15 and was obtained as in Ngo et al. (2004), whereas the number of nodes in the low-rank formulation of the spatial component was obtained by CLARA algorithm (Kaufman et al., 1990). The coordinates of the 23 spatial nodes were obtained by a space-filling algorithm implemented in function `default.knots.2D` within the R library SemiPar. The low rank formulation of model 1 was estimated by REML using function `lme` of the R library `nlme` (Pinheiro and Bates, 2000). The three cross-validation techniques CV_1, CV_2 and CV_3 suggested by Carroll and Cressie (1996) were used to compare the accuracy and the precision of estimates of the two models. In particular CV_1 was used to assess the unbiasedness of the predictor (optimal value: CV_1=0), CV_2 was used to assess the accuracy of the mean squared prediction error (optimal value: CV_2=1) and CV_3 was used to check the goodness of prediction (small value of CV_3 indicates a good fit).

3. Results

The result comparison suggests that the two fitted models have good and similar performances and are very useful for analyzing the relationship between *HI* and the covariates during the two crop years (Table 1). For this reason the most generally used exponential covariance structure was chosen. The fitted geoadditive model, obtained by using the exponential correlation structure to specify the spatial dependence of the geographical component, is reported in Table 2. From the table inspection, one sees that both agronomical variables (*EW*, *FP*) impact significantly on *HI*, however the relationship with *FP* is more complex, due also to the higher uncertainty in *FP* measurement. All nonlinear components of the geoadditive model are significant on the basis of the degrees of freedom (Table 2) estimates that confirmed the appropriateness of including the nonlinear effects of *FP*, the spatial component and the interaction between the factor year and the overall spatial effect in the fitted model.

Spatial correlation structure	CV_1	CV_2	CV_3
Exponential	-0.61	1.29	8.44
Gaussian	-0.62	1.31	8.42

4. Concluding remarks

The proposed approach is a quick and effective method of predicting the spatial distribution of the harvest index using standard agronomic measurements over two years. The great advantages of geoadditive models lie mainly in the possibility to jointly analyse spatial and temporal variations and to treat the complex interactions, quite often non linear, between production process and several different variables (soil, crop, atmosphere, management). Moreover, these models allow us to predict agronomical variables in specific locations of the field and this piece of information is crucial for Precision Agriculture. These considerations and the possibility of estimating linear

effects and variance components of non linear effects and error term by REML, using mixed effects model procedures routinely implemented in statistical software, lead us to recommend a wider use of geoadditive models with interactions in the presence of spatial dependence and temporal variation.

Table 1: Cross-validation errors with two different spatial correlation structures

Linear component			
Covariates	**Coefficients**	**Std.Error**	**p-value**
EW	0.092	0.028	**<0.05**
Non-linear component			
Covariates	**df**	**N° knots**	
FP	**9.35**	**15**	
***locations of year*$_{2006}$**	**8.02**	**15**	
***locations of year*$_{2008}$**	**8.02**	**15**	
X, Y	**7.02**	**23**	

Table 2: Summary of the REML based fit of the model with exponential correlation structure.

References

Cafarelli B., Castrignanò A. (2011). The use of geoadditive models to estimate the spatial distribution of grain weight in an agronomic field: a comparison with kriging with external drift, *Environmetrics* 22.

Carroll SS, Cressie N. (1996). A comparison of geostatistical methodologies used to estimate snow water equivalent. *Water Resour. Bull.* 32: 267– 278.

Greven S, Crainiceanu CM, Kuechenhoff H, Peters A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, 17: 870-891

Kammann EE, Wand MP. (2003). Geoadditive models, *Applied Statistics*, 52: 1-18.

Minasny B. and McBratney A.B. (2005). The Matérn function as a general model for soil variograms, Geoderma 128, 192-207.

Pinheiro JC and Bates DM. (2000). *Mixed-Effects Models in S and S-Plus*, Springer Verlag: New York

Ruppert D, Wand MP, Carroll RJ. (2009). Semiparametric Regression During 2003-2007. *Electronic Journal of Statistics*, 3: 1193-1256

Stein ML. 1999. *Interpolation of Spatial Data – Some Theory for Kriging*. Springer Verlag, Ney York.

White E.M., Wilson F.E.A. (2006). Responses of grain yield, biomass and harvest index and their rates of genetic progress to nitrogen availability in ten winter wheat varieties. *Irish Journal of Agricultural and Food Research* **45**: 85–101.

On the design-based properties of spatial interpolation[1]

Francesca Bruno, Daniela Cocchi, Alessandro Vagheggini
Dipartimento di Scienze Statistiche, Via Belle Arti 41 Bologna,
{francesca.bruno; daniela.cocchi; alessandr.vagheggin2}@unibo.it

Abstract: When spatial interpolation is carried out under a deterministic approach rather than according to the classical model-based approach known as kriging, the statistical properties of the predictor cannot be assessed. The aim of this work is to achieve these properties under a finite population design-based framework, that treats spatial locations as the outcome of a probabilistic sample.

Keywords: spatial sampling; ratio estimator, design based inference; spatial information in finite population inference.

1. Introduction

Given n locations $\mathbf{u}_1,\ldots,\mathbf{u}_n$ over a surface, let us consider a fixed but unknown deterministic function $z(\cdot)$ which generates the data $z(\mathbf{u}_1),\ldots,z(\mathbf{u}_n)$. The inverse distance weighted interpolator (IDW, Shepard, 1968) for predicting the value in an unknown location (denoted by a Greek letter) is

$$\hat{\zeta}(\mathbf{u}_{\bar{\lambda}}) = \mathbf{z}'\mathbf{w}_{\bar{\lambda}} \quad , (1)$$

where the normalized inverse squared distances of the unknown location from all the sampled ones $w_i = \left\|\mathbf{u}_{\bar{\lambda}} - \mathbf{u}_i\right\|^{-2} \Big/ \sum_{j=1}^{n} \left\|\mathbf{u}_{\bar{\lambda}} - \mathbf{u}_j\right\|^{-2}$ are contained in the weighting vector $\mathbf{w}_{\bar{\lambda}} = (w_1,...,w_i,...,w_n)'$ and $\mathbf{z}$ is the n-dimensional vector of the observed values. The IDW properties are well known; the predictor conforms to the Tobler's law of geography. Here we propose to view this predictor under a design-based perspective.
Let us now consider the n locations as a probabilistic sample from a population of N (Barabesi, 2008): the unknown values at the unsampled locations are the object of the inference.

2. The Inverse Distance Weighted interpolator in the finite population framework

Under the design-based framework, the IDW interpolator can be seen as the result of a sampling procedure. Since each individual unobserved value depends only on its unique specific geographical relationship with the sampled locations, the simple random sampling without replacement is chosen.

[1] Work supported by the project PRIN 2008: New developments in sampling theory and practice, Project number 2008CEFF37, Sector: Economics and Statistics, awarded by the Italian Government.

The sampling design can be suitably taken into account through the use of random selection matrices (Bruno *et al*., 2011), that allow to pass from sample-based quantities to population-based ones. Expression (1) becomes

$$\hat{\zeta}(\mathbf{u}_{\bar{\lambda}}) = \mathbf{z}'\mathbf{w}_{\bar{\lambda}} = \frac{\boldsymbol{\zeta}'\mathbf{A}_{\bar{\lambda}}\boldsymbol{\Phi}\mathbf{b}_{\bar{\lambda}}}{\mathbf{1}'\mathbf{A}_{\bar{\lambda}}\boldsymbol{\Phi}\mathbf{b}_{\bar{\lambda}}} = \frac{\boldsymbol{\zeta}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}}{\mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}} \quad , (2)$$

where $\boldsymbol{\Phi}$ is a $N \times N$ symmetric matrix containing the same function of the Euclidean distances $\left\|\mathbf{u}_{\bar{\lambda}} - \mathbf{u}_{\lambda}\right\|^{-2}$ of (1) before normalization with null diagonal, while $\boldsymbol{\varphi}_{\bar{\lambda}}$ is its $\bar{\lambda}$-th column vector ($\bar{\lambda} = 1,\ldots,N$). $\mathbf{A}_{\bar{\lambda}}$ is the diagonal matrix containing the random conditional indicator variables $I_{(\lambda \in s | \bar{\lambda} \notin s)}$ and $\mathbf{b}_{\bar{\lambda}}$ is the randomization of the $\bar{\lambda}$-th canonical basis vector $\mathbf{e}_{\bar{\lambda}}$ through the random indicator variable $I_{(\bar{\lambda} \notin s)}$. Using some matrix algebra and results for conditional random variables one can see that $\mathbf{K}_{\bar{\lambda}}$ is the diagonal matrix of the joint random indicator variables $I_{(\bar{\lambda} \notin s, \lambda \in s)}$. The IDW interpolator is written in (2) as a function of the *N*-dimensional vector of population values $\boldsymbol{\zeta}$ and of random indicator variables. Through the use of selection matrices, sampled and unsampled locations are associated in order to manage exclusion and conditional inclusion in the sample through random indicator variables. The resulting predictor turns out to be a design-based ratio-type estimator (Särndal *et al*., 1992).

3. Approximated first two moments of the IDW interpolator

Rewriting the IDW interpolator as in (2) allows the calculus of its statistical properties. Since it is a ratio of linear random combinations, its properties can be analytically computed only as approximations. For managing the involved random variables, we define the "association probabilities", linking a potentially unsampled location with all the others. These probabilities represent the starting point for the calculus of the statistical properties of the predictor: an uncertainty measure can, in this way, be associated to the deterministic IDW interpolator.

Theorem 1: The approximated expected value of (2) is

$$\mathrm{E}\left[\hat{\zeta}(\mathbf{u}_{\bar{\lambda}})\right] = \mathrm{E}\left[\frac{\boldsymbol{\zeta}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}}{\mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}}\right] \cong \frac{\mathrm{E}\left[\boldsymbol{\zeta}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]}{\mathrm{E}\left[\mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]} = \frac{\sum_{\lambda \neq \bar{\lambda}} \zeta_{\lambda}\varphi_{\lambda\bar{\lambda}}}{\sum_{\lambda \neq \bar{\lambda}} \varphi_{\lambda\bar{\lambda}}} = \frac{T_{1\bar{\lambda}}}{T_{2\bar{\lambda}}} \quad (3)$$

Proof.
It follows directly from the expected value of the random matrix $\mathbf{K}_{\bar{\lambda}}$ as

$$\mathrm{E}[\mathbf{K}_{\bar{\lambda}}] = \frac{N-n}{N}\frac{n}{N-1}\mathbf{D}_{\bar{\lambda}} \quad , (4)$$

where $\mathbf{D}_{\bar{\lambda}}$ is a diagonal matrix of unit values besides the null value at the $(\bar{\lambda},\bar{\lambda})$-th position. □

Let us define the difference between each $\bar{\lambda}$-th population value and its interpolation via the other $N-1$ values

$$\delta(\mathbf{u}_{\bar{\lambda}}) = \zeta(\mathbf{u}_{\bar{\lambda}}) - \left(\sum_{\lambda \neq \bar{\lambda}} \zeta_{\lambda} \varphi_{\lambda\bar{\lambda}} \Big/ \sum_{\lambda \neq \bar{\lambda}} \varphi_{\lambda\bar{\lambda}} \right), \quad (5)$$

as the "structural bias" associated to location $\mathbf{u}_{\bar{\lambda}}$. The bias of estimator (2), *i.e.* $\mathrm{E}\left[\hat{\zeta}(\mathbf{u}_{\bar{\lambda}})\right] - \zeta(\mathbf{u}_{\bar{\lambda}})$, is also not null. However, it can be seen that, as the sample size increases, $\hat{\zeta}(\mathbf{u}_{\bar{\lambda}})$ tends to its "true" value $T_{1\bar{\lambda}}/T_{2\bar{\lambda}}$ (3). Predictor (2) may exhibit a high "structural bias" due not to the sample size but to the nature of the interpolator.

Theorem 2. The approximated variance of (2) is

$$\mathrm{V}\left[\hat{\zeta}(\mathbf{u}_{\bar{\lambda}})\right] \approx \frac{\mathrm{V}\left[\boldsymbol{\zeta}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]}{\mathrm{E}\left[\mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]^2} - 2\frac{\mathrm{Cov}\left(\boldsymbol{\zeta}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}, \mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right)}{\mathrm{E}\left[\mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]^3} + \frac{\mathrm{E}\left[\boldsymbol{\zeta}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]^2 \mathrm{V}\left[\mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]}{\mathrm{E}\left[\mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]^4},$$

which, using a notation similar to (3), can be expressed as

$$\mathrm{V}\left[\hat{\zeta}(\mathbf{u}_{\bar{\lambda}})\right] \approx \frac{1}{cT_{2\bar{\lambda}}^4}\left[h\left(T_{3\bar{\lambda}}T_{6\bar{\lambda}} - 2T_{7\bar{\lambda}}T_{8\bar{\lambda}} + T_{5\bar{\lambda}}T_{1\bar{\lambda}}^2\right) + m\left(T_{4\bar{\lambda}}T_{6\bar{\lambda}} - 2T_{8\bar{\lambda}}^2 + T_{6\bar{\lambda}}T_{1\bar{\lambda}}^2\right)\right],$$

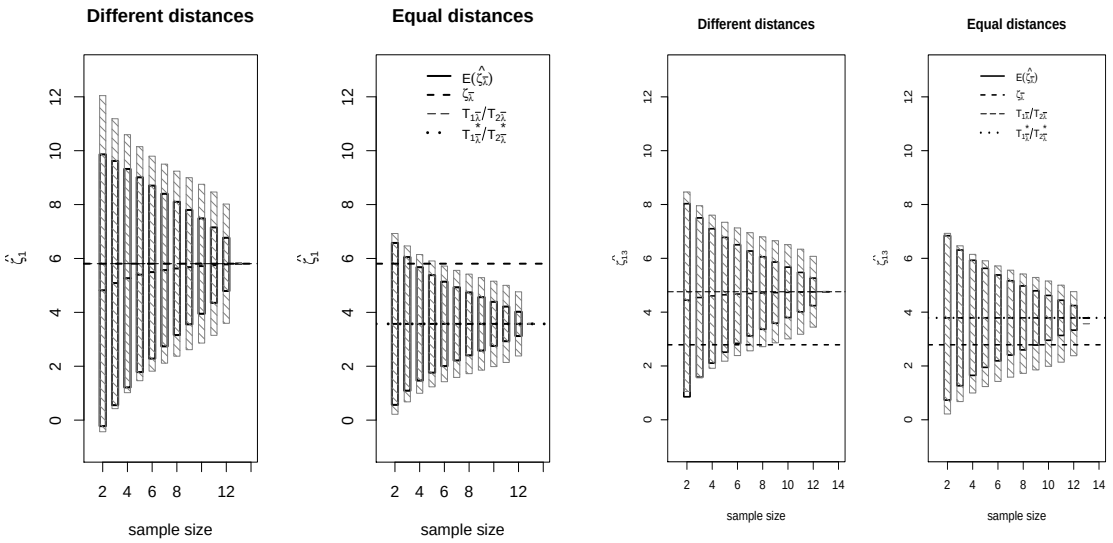
where c, h and m are population constants and quantities $T_{\square\bar{\lambda}}$ are similar to those in (3). For the proof, see Bruno *et al.* (2011). □

4. A simulation study

We assess the improvement in inference provided by the use of a weighting system based only on geographical distances. No model specification is required and the only assumption made is that data follow the Tobler's law. The weighting system we propose, suggested by the IDW interpolator (1), is the same for the whole population, but the weights change according to the location to predict. When geography is not important, it might be more useful to predict the unweighted mean of the *N*-1 population values, for the unknown location.

A simulation study has been carried out for evaluating the approximate properties of the IDW interpolator under the design-based framework. A population of fifteen sparse data points is considered. A map of the population under study, the table of the values of the variable and the "structural bias" associated to each point of the population are given in Bruno *et al.* (2011). We illustrate two opposite situations, in the four panels of Figure 1. For the first location, where $\zeta(\mathbf{u}_1) = 5.81$, the structural bias is null: the use of the distances, linked to the IDW predictor, leads to a better prediction (panel a) than the consideration of equal weights (panel b), as highlighted by the tendency of the expected value (3) to the real value. The other location, where $\zeta(\mathbf{u}_{13}) = 2.79$, presents a structural

bias $\delta(\mathbf{u}_{13}) = -1.97$ and (3) fails in properly predicting the true value (panel c). The unweighted version of the structural bias is on the contrary $\delta^*(\mathbf{u}_{13}) = -1.00$. For this point, the use of geography is misleading and a situation of unweighted estimation in simple random sampling would be preferable (panel d).



a) b) c) d)

Figure 1: Prediction with different and equal weights for two locations (as *n* increases). Location 1: panels a) and b); location 13: panels c) and d).

References

Barabesi L. (2008) Facoltà di Economia "R.M. Goodwin", Università degli Studi di Siena, mimeo.

Bruno, F., Cocchi, D. and Vagheggini, A. (2011) Spatial interpolation using a finite population approach, submitted.

Cressie N.A.C. (1993) *Statistics for spatial data*, Wiley, New York.

Shepard D. (1968) A two-dimensional interpolation function for irregularly-spaced data, *Proceedings of the 1968 23rd ACM national conference*, 517-524.

Särndal C.-E., Swensson B., Wretman J. (1992) *Model-assisted survey sampling*, Springer-Verlag, New York.

Stevens D.L. (2006) Spatial properties of design-based versus model-based approaches to environmental sampling, *American Statistical Association; Section on Statistics & the Environment Newsletter*, 10, 3-5.

Ver Hoef J.M. (2002) Sampling and geostatistics for spatial data, *Ecoscience*, 9, 152-161.

Relations between spatial design criteria [1]

Werner G. Müller and Helmut Waldl
Department of Applied Statistics, Johannes-Kepler-University Linz,
werner.mueller@jku.at

Abstract: Several papers have recently strengthened the bridge connecting geostatistics and spatial econometrics. For these two fields various criteria have been developed for constructing optimal spatial sampling designs. We will explore relationships between these types of criteria as well as elude to space-filling or not space-filling properties.

Keywords: Empirical kriging, compound D-optimality, Moran's I

1 Introduction

Lindgren et al. (2011) further strengthen the bridge connecting the two somewhat disparate worlds of spatial analysis. One is rooted in the idea of observing continuously varying spatial processes and led to what is largely referred to as geostatistics. The other, which assumes (usually aggregate) observations attached to discrete (mostly irregular) lattices, is commonly known under the name of spatial econometrics. In particular in the latter literature the rift between these two points of view - manifesting itself along various themes - is a constant challenge towards a unified understanding (Griffith and Paelinck, 2007). Also for the more narrow topic of efficient estimation and prediction early contributions can be found there (Griffith and Csillag, 1993) and that the issue is of great current interest is documentable as well (Fernández-Avilés Calderón, 2009). The method of explicitely linking some Gaussian fields to Gaussian Markov random fields on irregular grids given in Lindgren et al. (2011) is certainly a very welcome addition to the equipment connecting the two views as the authors rightfully claim in their discussion section. It remains to be seen whether practitioners will be able to take it up as easily as a perhaps more pragmatic recent suggestion like Nagle et al. (2011).

2 Materials and Methods

But let us draw the attention towards a rather neglected (in the discussion section of Lindgren et al. (2011) as well most of the literature in general) aspect of establishing such a link as above. That is the potential impact of this link on the respective optimal sampling designs and the question of their effective generation.

[1] A considerably shortened and edited version of this paper will be published as a discussion of Lindgren et al. (2011) in JRSS-B.

We will illustrate our points on the same example as used in Section 2.3 of Lindgren et al. (2011), namely the leukaemia survival data, utilizing some of the calculations thankfully provided by the authors.

In geostatistics the optimal sampling design is often based upon the kriging variance over the region of interest $\mathcal{X}$, frequently by minimizing its maximum. It has turned out that this reflects rather not so well the true variation as the uncertainty introduced by estimating covariance parameters γ is thereby neglected. To compensate for that Zhu and Stein (2006) and Zimmerman (2006) have suggested minimizing the modification

$$\max_{x \in \mathcal{X}} \left\{ \mathrm{Var}[\hat{Y}(x)] + \mathrm{tr} \left\{ M_\gamma^{-1} \mathrm{Var}[\partial \hat{Y}(x)/\partial \gamma] \right\} \right\},$$

which the latter has termed the EK(empirical kriging)-criterion. Here M_γ stands for the Fisher information matrix with respect to γ, and we can analogously denote M_β for trend parameters β for later usage.

In spatial econometrics it is common to test for spatial autocorrelation by specifying a spatial linkage or weight matrix W and utilize an overall type measure such as Moran's I. Therefore Gumprecht et al. (2009) have suggested to employ the power of Moran's I under a hypothesized spatial lattice process given by its precision matrix Q as the design criterion; let us call maximization of it the MIP(Moran's I power)-criterion in the following.

3 Results

Now as there is a link established with respect to estimation between the two modelling paradigms, can we expect a similar link with respect to those associated design criteria? Looking at the example a sensible design question we could pose is to which out of the 24 districts in north-west England should we sample if we are limited to a number $k < 24$ for financial reasons. To keep things simple, we will in the following choose $k = 3$, which allows for $\binom{24}{3} = 2024$ different designs. For all those designs we can then calculate the values for the above design criteria and plot them against each other to judge for a potential linkage. As the only covariance parameter, which is not predetermined in the example is ρ, we have $\gamma = \rho$ and EK reduces to scalar operations localized at $\rho = 0.2$. For the MIP we required the precision matrix Q, which was provided by Lindgren et al. (2011). The matrix W was defined by assigning 1 to point pairs with intersite distances less than the range $\rho = 0.2$ and 0 else, which turned out to be an insensitive choice.

At this point we now had to slightly modify the example: since the spatial correlation is so strong in the leukaemia data most of the realized powers were very close to one, thus obscuring all potential patterns. We therefore artifically reduced the number of cases (and thus the powers) by randomly sampling 20 locations from the 3 districts respectively. This resulted in the scatter plot of criteria displayed in the left panel of the figure in the discussion of Lindgren et al. (2011). While

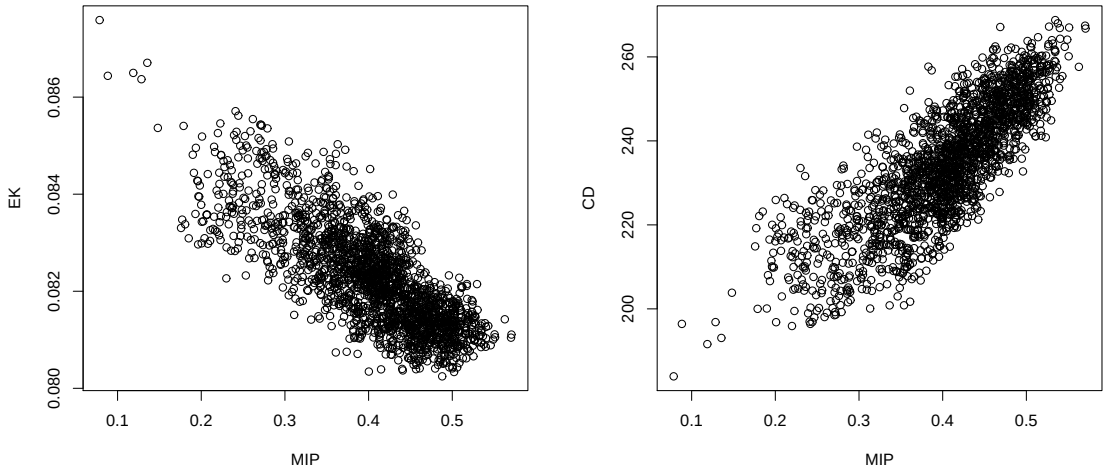


Figure 1: left panel: MIP (horizontal) versus EK (vertical) criterion values; right panel: MIP (horizontal) versus CD (vertical) criterion values.

from this display the link between the criteria already becomes quite evident, we present here in Figure 1 an even stronger one well extending into the corners where the optima lie. This was achieved by simply doubling the diagonal entries of the covariance matrix Q^{-1}, which emulates a stronger nugget effect.

It thus looks that in cases with reasonable localized spatial dependence one could achieve reasonably high design efficiencies by employing one for the other criterion, which offers advantages in both directions. Where MIP requires little prior knowledge its optimization is nonstandard, whereas for EK and related cases well developed theory is available (Müller and Pázman, 2003).

Both criteria, however, are computationally quite intensive and it makes thus sense to look for cheaper alternatives. Motivated by the traditional connection between estimation and prediction based criteria ("equivalence theory"), Müller and Stehlík (2010) have suggested to replace the EK-criterion by a compound criterion for determinants of information matrices, i.e. maximizing

$$|M_\beta)|^\alpha \cdot |M_\gamma|^{(1-\alpha)},$$

with a weighing factor α, which we will call in the following CD_α(compound D)-optimality. The relationship of this criterion (assuming a constant trend β) with an $\alpha = 0.5$ to the MIP is displayed in the right panel of Figure 1. This clearly shows that one could computationally very cheaply find the optimum with respect to CD and still achieve rather high efficiencies on the MIP criterion.

4 Concluding remarks

We must note that our calculations have shown that the dependence between the criteria is related to the specific setup. It turns out that the strength of the relation-

ship between MIP and the other two criteria decreases when the powers approach one, but strongly increases for decreasing ranges and increasing nuggets. Note also the relationships to the ubiquitous space-filling designs as explored in Pronzato and Müller (2011). Summarizing, we believe our discussion showed that the relations between the two linked approaches can go far beyond mere estimation issues.

References

Fernández-Avilés Calderón, G. (2009). Spatial regression analysis vs. kriging methods for spatial estimation. *International Advances in Economic Research 15*(1), 44–58.

Griffith, D. and J. Paelinck (2007). An equation by any other name is still the same: on spatial econometrics and spatial statistics. *The Annals of Regional Science 41*(1), 209–227.

Griffith, D. A. and F. Csillag (1993). Exploring relationships between semivariogram and spatial autoregressive models. *Papers in Regional Science 72*(3), 283–295.

Gumprecht, D., W. G. Müller, and J. M. Rodríguez-Díaz (2009). Designs for detecting spatial dependence. *Geographical Analysis 41*(2), 127–143.

Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach *Journal of the Royal Statistal Society Series B 73*(4), forthcoming.

Müller, W. G. and A. Pázman (2003). Measures for designs in experiments with correlated errors. *Biometrika 90*(2), 423–434.

Müller, W. G. and M. Stehlík (2010). Compound optimal spatial designs. *Environmetrics 21*(3-4), 354–364.

Nagle, N. N., S. H. Sweeney, and P. C. Kyriakidis (2011). A geostatistical linear regression model for small area data. *Geographical Analysis 43*(1), 38–60.

Pronzato, L. and W. G. Müller (2011). Design of computer experiments: space filling and beyond. *Statistics and Computing*, Online First.

Zhu, Z. and M. L. Stein (2006). Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics 11*(1), 24–44.

Zimmerman, D. L. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics 17*(6), 635–652.

Simulation-based optimal design for estimating weed density in agricultural fields

Bel Liliane, Parent Eric
AgroParisTech/ INRA MIA 518, Paris, France. Liliane.Bel@agroparistech.fr

Makowski David
INRA, Thiverval-Grignon, France.

Abstract: In order to manage herbicide treatment we present a method for optimizing the locations of weed density measurements. The practical problem is to estimate weed density in each one of the n quadrats of a field, assuming that m measurements were already collected and using p additional measurements optimally located. The proposed method consists in three steps: 1) fit a statistical model to the m available measurements taking into account the nature of the data, 2) define possible locations of the p additional measurements using a simulated-annealing algorithm, 3) assess the designs using weed density values simulated using the fitted statistical model. This method is applied to several wheat fields and the results show that it improves weed density predictions. Sensitivity to several tuning parameters is discussed.

Keywords: optimal design, spatial statistics, weed

1 Introduction

Weeds can induce important yield losses in agricultural fields. In order to prevent huge losses weed management is frequently based on herbicide application. But extensive herbicide application leads to a risk of water pollution by chemicals. Sometimes, herbicide application is useless and the need of precise knowledge of weed density in the field is crucial. In order to provide a map of weed density in a field without counting all the plants, it is necessary to design a spatial statistical model fitted from a limited number of measurements. The purpose of this paper is to present a method for optimizing the locations of weed density measurements in agricultural fields in order to manage herbicide treatment. Consider an agricultural field divided into n quadrats and assume that weed density measurements were already collected in m out of the n quadrats, $m < n$. Our practical problem is to estimate weed density in each one of n quadrats by using

i) the m available weed density measurements,

ii) p additional measurements, $p < n - m$, collected in other quadrats located in the same field,

and by estimating the weed density in the unmeasured quadrats with a statistical technique. Potentially, the use of p additional measurements can lead to improved weed density estimates, but the degree of improvement depends on the experimental design i.e. on the number of additional measurements p and on their locations in the field.
This paper presents a method for defining, assessing, and selecting experimental designs in order to determine an appropriate number p of additional measurements and optimize their locations in the field.
The proposed method consists in three main steps:

1. fit a spatial statistical model to the m available measurements taking into account that the data are countings or presence-absence data,
2. assess the design of $m + p$ quadrats using weed countings values simulated using the fitted statistical model to define the criterion,
3. define possible locations of the p additional measurements using a simulated-annealing algorithm according to the previously defined criterion.

This method was applied to several wheat fields and the results showed that it could improve weed density predictions. Sensitivity to several tuning parameters is discussed.

2 Materials and Methods

2.1 Statistical model for mapping weeds

Assuming m measurements are available, a standard technique to produce a map of weed countings in a field of n quadrats is ordinary kriging. Kriging performs well when the data distribution is Gaussian or not far from Gaussian. Weed countings are discrete data and the Gaussian distribution is not well adapted. Models for Poisson, zero Inflated Poisson and binary data are designed involving a continuous Gaussian latent variable accounting for the spatial dependence. The kriging is performed on the latent variable.

2.2 Conditional simulations

Conditional simulations are simulations of a spatial field according to a spatial model which are constrained to take observed values in a set of locations. Given a design of $m + p$ sites, its quality is assessed with the root mean square difference between conditional simulations and kriging estimates.

2.3 Simulated Annealing Algorithm

The search of an optimal design is achieved by a simulated annealing algorithm: p quadrats are randomly selected and added to the m initial quadrats. Slight perturbations on the previous configuration are iteratively proposed to improve the

conditional simulation criterion. Configurations that do not improve the criterion are accepted with a decreasing probability in order to favour the exploration of the configurations domain.

3 Results

3.1 Simulated data

The procedure is evaluated on simulated data, sharing the same characteristics as the weed data (size of the field, number of quadrats, countings data of same magnitude). It turns out that the procedure gives a better design to estimate a weed map when the data are significantly spatially correlated provided that the variogram of the latent variable is well estimated. Not surprisingly when the data are not or slightly correlated a random design does the job as well. As usual the simulated annealing algorithm is sensitive to the temperature parameter that has to be tuned accordingly to the magnitude and the kind of the data (countings or presence-absence). Several ways to modify the design configuration (all the p points or only one are randomly changed, the modification is random on one or two directions), have been tested but they result in equivalent outcomes.

3.2 Case study

Weeds have been measured exhaustively in a field divided in 92 quadrats on a grid 4×23. $m = 20$ regularly arranged sites are selected in such a way to cover the entire domain. We look for $p = 10$ other points to improve the estimate of the weed map. The procedure is achieved with 1000 iterations for the SAA. Figure 1 shows a) the original data and the 20 points of the initial design, b) the estimated map with 10 sites randomly selected added to the initial design and c) the final design with the optimized 10 additional sites. The RMSE has improved by 15%, and it is worth noticing that the procedure locates the new sites in the area where the weeds are numerous.

4 Concluding remarks

Accuracy of predicted infestation levels depends on locations of weed density measurements. We showed that locations leading to accurate predictions can be found using a simulation-based approach with a simulated-annealing step. This approach can be used to map weed infestation in agricultural fields and allows farmers to apply herbicides in highly infested areas only. The performance of the proposed approach depends on the spatial correlation of weed densities and on tuning parameters of the simulated annealing algorithm. An algorithm based on particle filter could also be used within the same framework.

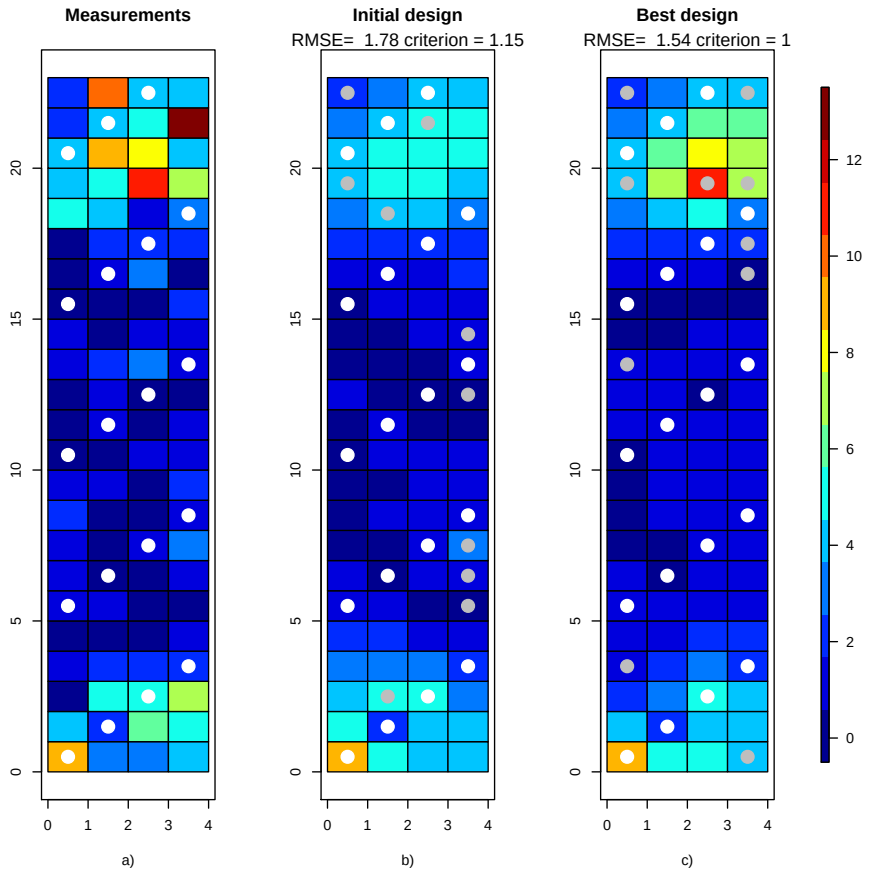


Figure 1: a) original data and initial design b) estimated map with 10 randomly additional sites c) estimated map with 10 optimized additional sites

References

Amzal B., Bois F.Y., Parent E. (2006). Bayesian-optimal design via interacting particle filter. *Journal of the American Statistical Association* 101, 773-785.

Makowski D., Chauvel B., Munier-Jolain N. (2010). Improving weed population model using a sequential Monte Carlo method. *Weed Research*, vol. 50, no4, pp. 373-382.

Müller P. (1999). Simulation-based optimal design. *Bayesian statistics* 6, 459-474. J.M. Bernado, J.O. Berger, A.P. Dawid, A.F.M. Smith (Eds.). Oxford University Press.

The dramatic effect of preferential sampling of spatial data on variance estimates [1]

David Clifford, Petra Kuhnert, Melissa Dobbie, Jeff Baldock, Neil McKenzie, Bronwyn Harch
CSIRO, David.Clifford@csiro.au

Ichsani Wheeler, Alex McBratney
University of Sydney

Abstract: Classic probability-based designs are widely used for spatial sampling in environmental research. When sampling over large regions researchers may wish to preferentially sample some sites due to ease of access. If such non-standard probability designs are implemented, Horvitz-Thompson analysis provides unbiased estimates for spatial means and variances provided first and second order inclusion probabilities can be evaluated. However, even with minor departures from standard designs the effect of preferential sampling on the sampling variance can be dramatic. We find significant increases in sampling variance as sampling becomes more and more preferential. We conclude that some non-standard designs can result in significantly weaker sampling performance and recommend they be examined by simulation prior to implementation.

Keywords: Probability design, GRTS, Horvitz-Thompson Estimators

1. Introduction

There are two broad categories of approaches available for surveying soil organic carbon (SOC) across space - model-based approaches and design-based approaches. The former set of approaches is very useful for mapping and prediction but is based on strong assumptions on the distributional properties of SOC. The latter is based entirely on how sites are chosen for sampling and can produce unbiased estimates of mean SOC as well as unbiased estimates of sampling variance. We examine design-based approaches here as they have been garnering more and more attention in recent years.

Probability-based designs have not been implemented on a national scale within Australia. A major challenge to establishing a national monitoring scheme is the large distances one would need to travel to collect data. For many regional sampling schemes there is anecdotal evidence that sample sites tend to be “just inside the gate, along the fence 50m from the road” which indicates a preference for sites that are easy to access. This is a defining feature of what we term the Australian context and this feature can bias the results of an otherwise well designed experiment when the true manner in which sites are chosen is not incorporated into the analysis.

In this report we explore designs that are compatible with the Australian context, i.e. designs that preferentially sample sites that are easy to access over remote sites. Using classic statistical design methodology coupled with modern computer simulation

[1] Financial support for this research comes from the CSIRO Sustainable Agriculture Flagship.

strategies we explore the effects of such preferential sampling on sampling variance. While stratified sampling generally improves sampling variance relative to simple random sampling, preferential sampling negates this benefit. However, we find that the modern technique of generalised random tessellation stratification (GRTS) sampling can incorporate preferential sampling quite well. In all our examples preferential sampling leads to increases in sampling variance but for GRTS this increase is not fatal

2. Preferential Sampling Probability Designs

We compare the performance of simple random sampling (SRS), stratified random sampling (STR) and GRTS using two spatial datasets. Cochran (1977) provides a detailed summary of many classic sampling designs and analysis results including the work of Horvitz and Thompson (1952) for computing unbiased estimators of mean and sampling variance using first and second order inclusion probabilities. GRTS was developed for sampling streams and stream networks (Stevens and Olsen 2003) and can readily handle any set of first order inclusion probabilities. GRTS has been used extensively in the U.S. by the Environmental Protection Agency for water-based monitoring (e.g. Schweiger et al 2005, Wardrop et al 2007) and can also be used for monitoring natural resources in terrestrial applications (Fancy, Gross and Carter 2009) though to the best of our knowledge it has not been used for soil carbon monitoring.

We venture away from classic designs by specifying inclusion probabilities in a manner that preferentially samples sites that are closer to roads that span the space of interest. We parameterise a linear relationship between inclusion probability and distance to road using a single term α that ranges from 0 to 1. When $\alpha = 0$ the linear relationship is flat, i.e. all inclusion probabilities are equal and we have classical non-preferential sampling. When $\alpha=1$ the inclusion probabilities for the sites furthest from the roads are zero. This boundary case is not considered since a design-based approach is no longer applicable to the whole region of interest. For values of α between 0 and 1 the inclusion probabilities decrease with distance, and the rate of decrease increases with α.

We use the work of Hartley and Rao (1962) to sample a specific number of sites according to our pre-specified first order inclusion probabilities as well as for computing approximations for our second order inclusion probabilities, simplified further by Stehman and Overton (1994). These can be used to compute Horvitz-Thompson estimators from implementations of non-standard designs.

3. Data

We use two spatial datasets to evaluate these probability designs. The first is a simulated non-stationary, non-isotropic process from fixed rank kriging of a spatial random effects model (Cressie & Johannesson, 2008). Values for this process are evaluated at 4 million pixels and we draw samples of size n=27. A grid of nine square strata is used for STR for this dataset. The second is a dataset of over 2.5 million predictions of percentage SOC across a large part (150,000 squared-km) of New South Wales in Australia (Wheeler et al, 2010). These predictions come from a Cubist-based data-mining model of legacy %SOC data from the Australian Soil Resource Information System (ASRIS, McKenzie et al 2005). We draw samples of size n=150 from the SOC

dataset. We define 16 strata for this dataset based around the major towns of the region with each site allocated to the stratum associated with the closest town.

4. Methods

We repeatedly apply the probability designs to our two datasets changing the strength of preferential sampling through the parameter α. For each design and α value we examine the distribution of our estimates of sampling variance. Effective sample sizes are found by matching the median sampling variance with sampling variance estimates based on non-preferential SRS. As is well known, when sampling a spatial process, switching from SRS to STR or GRTS leads to an immediate large jump in effective sample size. We wish to investigate what happens to sampling variance and effective sample size as α changes from 0 to just under 1 for each design.

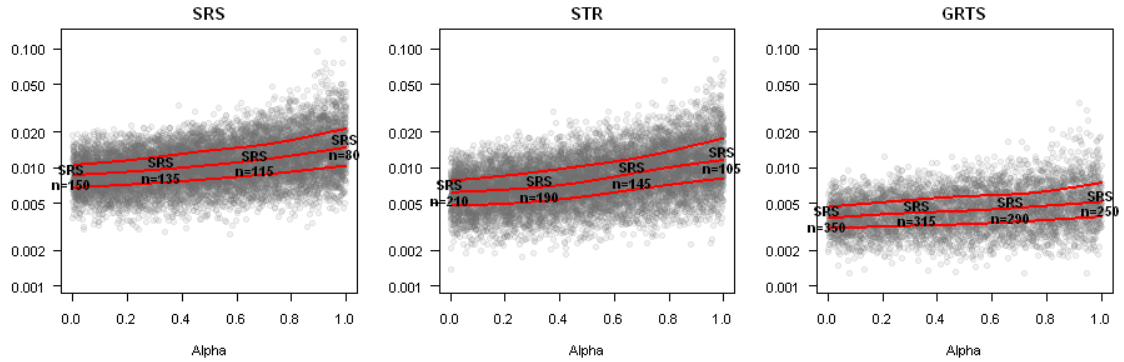


Figure 2: Effect of preferential sampling on the sampling variance of mean estimates for the SOC dataset under SRS, STR and GRTS designs. Red lines within each plot indicate 1st, 2nd and 3rd quartiles of the variance estimates. The text indicates selected *approximate* effective sample sizes under non preferential SRS designs based on smooth quantile regression.

5. Results and Discussion

Preferential sampling results in larger sampling variances in all cases. The gains in effective sample size one attains by switching to STR can be all but wiped out when preferential sampling is employed. Preferential sampling of n=27 sites under STR is routinely found to be worse than SRS based on far fewer sites. For GRTS the effect of preferential sampling is not as dramatic. Figure 2 plots our estimates of sampling variance for many values of α for each design for the SOC dataset. Each panel includes red lines based on smooth robust regression of the data to estimate the 1st, 2nd and 3rd quartiles as functions of α. The text written over each plot indicates effective sample sizes required to achieve similar sampling variances under non-preferential SRS.

This research indicates that continental-scale sampling schemes can be designed and implemented in a manner that better reflects how they are used in practice. While preferential sampling designs more accurately reflect practical concerns, we demonstrate that they can have dramatic inflationary effects on sampling variance. As such, we recommend a thorough evaluation of any sampling approach prior to implementation. In the examples explored here we found that estimates of sampling

variance from GRTS are least affected by preferential sampling. This suggests that GRTS is a viable approach for designing spatial sampling schemes at large scales. The success of GRTS is due partly to its use of a neighbourhood variance estimator (Stevens and Olsen 2004) and partly to the fact that GRTS achieves much better spatial balance compared to STR.

References

Cochran, W. G. (1977) *Sampling Techniques*, 3rd Edition Wiley

Cressie, N. & Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B*, 70, 209-226

Fancy, S.; Gross, J. & Carter, S. (2009) Monitoring the condition of natural resources in US national parks *Environmental Monitoring and Assessment*, 151, 161-174

Hartley, H. O. & Rao, J. N. K. (1962) Sampling with Unequal Probabilities and without Replacement *The Annals of Mathematical Statistics*, 33, 350-374

Horvitz, D. G. & Thompson, D. J. (1952) A Generalization of Sampling Without Replacement From a Finite Universe *Journal of the American Statistical Association*, 47, 663-685

McKenzie, N.; Jacquier, D.; Maschmedt, D.; Griffin, E. & Brough, D. (2005) Australian Soil Resource Information System (ASRIS) *National Committee on Soil and Terrain Information*

Schweiger, E.; Bolgrien, D.; Angradi, T. & Kelly, J. (2005) Environmental Monitoring and Assessment of a Great River Ecosystem: The Upper Missouri River Pilot *Environmental Monitoring and Assessment*, 103, 21-40

Stehman, S. V. & Overton, W. S. (1994) Comparison of Variance Estimators of the Horvitz-Thompson Estimator for Randomized Variable Probability Systematic Sampling *Journal of the American Statistical Association*, 89, 30-43

Stevens, D. L. & Olsen, A. R. (2003) Variance estimation for spatially balanced samples of environmental resources *Environmetrics*, 14, 593-610

Stevens, D. L. J. & Olsen, A. R. (2004) Spatially Balanced Sampling of Natural Resources *Journal of the American Statistical Association*, 99, 262-278

Wardrop, D.; Kentula, M.; Stevens, D.; Jensen, S. & Brooks, R. (2007) Assessment of wetland condition: An example from the Upper Juniata watershed in Pennsylvania, USA *Wetlands*, 27, 416-431

Wheeler, I.; Minasny, B.; McBratney, A. & Bui, E. (2010) A regional soil organic carbon prediction function for south-eastern Australia *19th World Congress of Soil Science*

Modeling malaria incidence in Sucre state, Venezuela using a Bayesian approach [1]

Desireé Villalta
Universidad Simón Bolívar, Caracas, Venezuela, villalta@cesma.usb.ve

Lelys Guenni
Universidad Simón Bolívar, Caracas, Venezuela

Yasmin Rubio
Universidad de Carabobo,Valencia, Venezuela

Abstract: This paper presents a hierarchical Bayesian Poisson lognormal model for malaria incidence in Sucre state, Venezuela, during the period $1990-2002$. The logarithm of the relative risk of the disease for each county or municipality is expressed as an additive model that includes a multiple regression with social-economic and climatic covariates; a random effect that captures the spatial heterogeneity in the study region and a CAR (Conditionally Autoregressive) component, that recognizes the effect of nearby municipalities in the transmission of the disease each year. For most years the selected model captures well the spatial structure between the relative risks from the nearby municipalities. When a poor model fit is obtained, a t-Student model for the spatial heterogeneity parameter improves model fitting results. From the 15 municipalities in Sucre state during the study period $1990-2002$, 7 of them presented high relative risks (greater than 1) in most years. These areas are mostly agricultural areas with poor living conditions.

Keywords: hierarchical Bayesian model, Poisson lognormal model, malaria incidence, Venezuela

1 Introduction

Malaria is a parasitic infectious tropical disease that causes high mortality rates in the tropical belt. In Venezuela, Sucre state is considered the third state with the highest malaria incidence. The *Standardized Mortality Ratio* (SMR), is the ratio between the number of observed disease cases (y_i) and the expected number of cases

[1]Project funded by the National Fund for Science and Technology (FONACIT) project No. 2005-000184, Venezuela.

in the region(E_i), this is, (Banerjee, 2003)

$$SMR_i = \widehat{\Psi}_i = \frac{y_i}{E_i} \qquad i = 1, ..., k \tag{1}$$

where k is the number of subregions (in our case the number of municipalities is 15) and $E_i = p^*.n_i = \frac{\sum_{i=1}^{k} y_i}{\sum_{i=1}^{k} n_i}.n_i$, being p^* the total proportion of disease incidence.

This incidence rate $\widehat{\Psi}_i$ is a raw estimate of the relative risk of disease infestation in the municipality i. A value grater than 1 indicates a disease incidence greater than expected for a region; therefore this constitutes an alarm for public health authorities, (Banerjee, 2003) and (Lawson, 2003). The objective of this work is to propose a model including temporal and spatial components, to explain the dynamics of the disease and to allow simultaneously to identify the explanatory social-economic and climatic variables related with the disease incidence in Sucre state.

2 Materials and Methods

2.1 Study region and Data

The study region is located in the northeastern region of Venezuela in Sucre state. This state has 15 municipalities with an area of $11,800 km^2$. Total cases of malaria were available for 13 years during the period $1990 - 2002$. Interpolated monthly precipitation was available for the whole state using a Bayesian Kriging approach (Le and Zidek, 2007). Several social-economic variables measuring basic needs coverage, unemployment rate, housing characteristics and public services were available from the National Institute of Statistics (INE). After a dimensional reduction technique based on principal component analysis (PCA), the following covariates were used from the PCA results:X_1: Percentage of households with fair building quality and lack of public services (electricity, sewerage, drinking water); X_2: Percentage of poor households with intermediate building quality; X_3: Sewerage and drinking availability; X_4: Percentage of population in agricultural activities. Additionally, the maximum monthly precipitation during the year, X_5, was also included. Each variable was stored in a matrix of dimension of 15×13.

2.2 Spatio-temporal model

Let Y_{it} the number of malaria cases in municipality i and year t. A Poisson model is usually assumed for these quantities, where the mean rate is $\lambda_{it} = E_{it}\, \Psi_{it}$. Therefore,

$$Y_{it} \sim Poisson(\lambda_{it}) \tag{2}$$

with $t = (1, ..., T)$, being T the number of years; in this case $T = 13$.

The proposed model for Ψ_{it} is:

$$\Psi_{it} = \exp(\alpha_t + \boldsymbol{\beta}_t.\mathbf{X}_{it} + v_{it} + b_{it}) \tag{3}$$

where $v_{it} \sim N(0, \frac{1}{\tau_{ht}})$ is a parameter representing the local spatial heterogeneity of the data and $b_{it}|b_{-it} \sim N\left(\bar{b}_{it}, \frac{1}{\tau_{bt} m_{it}}\right)$ is the Conditional Auto-Regressive (CAR) component representing the spatial dependence among the neighboring counties in the transmission of the disease. For model 3, we have the vectors $\alpha_t = (\alpha_1, \alpha_2, ..., \alpha_T)$, $\beta_t = (\beta_1, \beta_2, ..., \beta_T)$, $\tau_{ht} = (\tau_{h1}, \tau_{h2}, ..., \tau_{hT})$, $\tau_{bt} = (\tau_{b1}, \tau_{b2}, ..., \tau_{bT})$, $b_{it} = (b_{1t}, b_{2t}, ..., b_{kt})$, $v_{it} = (v_{1t}, v_{2t}, ..., v_{kt})$ and $\mathbf{X}_{it}$ is the covariates matrix.

As an alternative model, the spatial heterogeneity parameter v_{it} can also be assumed to have a $t - Student$ distribution. The complete conditional posterior probability distributions were calculated for parameters α_t, β_t, b_{it}, v_{it}, τ_{bt},τ_{ht}.

The prior distributions for the parameters α_t, β_t, v_{it}, b_{it}, τ_{ht}, τ_{bt} of model 3, were assumed as follows: α_t and β_t are assumed Uniformly distributed; $b_{it}|b_{-it} \sim N\left(\bar{b}_{it}, \frac{1}{\tau_{bt} m_{it}}\right)$; $\tau_{ht} \sim Gamma(a_h, d_h)$ and $\tau_{bt} \sim Gamma(a_c, d_c)$, where parameters $a_h = a_c = 0.5$, $d_h = d_c = 0.0005$; b_{-it} is the parameter vector without considering the municipality i at time t; and m_{it} are the neighbors to municipality i at time t; although the number of municipalities does not change with time, we use the above notation.

3 Results

A computer code in WinBUGS was implemented for Bayesian inference using MCMC methods. Fourteen thousand samples from the parameter posterior distributions were obtained and $4,000$ samples were used for burnin. Several models were proposed by using different sets of covariates and the lognormal models with and without the CAR component (b_{it}) were also compared. The Deviance Information Criteria (DIC) (Spielgelhalter et al., 2002), and the Minimum Posterior Expected Loss Criteria (D) (Gelfand and Ghosh, 1998) were used for model selection. The DIC criteria did not show important variations among models. The D critera was more sensitive to model variations and suggested that a model with a CAR component and variables X_1, X_2, X_4 and X_5 was more appropriate, since this model presented the lowest D value. Model residuals for the selected model were tested for independence by calculating the Moran's I posterior probability interval for all years.

Posterior predictive model checks were carried out by simulating $2,000$ replicates from the posterior predictive distribution for each municipality and each year. The posterior predictive p-value $p(y_{it}^{rep} \leq y_{it}^{obs})$ was calculated to compare the observed vs. simulated values. If the p-value is close to 0 or 1, it means that the observed values are very unlike to be seen from the simulated values.

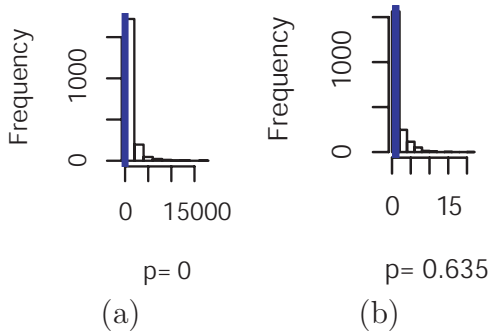


Figure 1: Posterior predictive check for the county Cruz Salmeron Acosta, year 1997 and calculated p-value, by using a normal model (a) and a $t - Student$ model (b) for the spatial heterogeneity parameter $v_{[i]}$

Model checks were satisfactory for most years and all municipalities, except for year 1997 with a good model fit only in 8 of 15 municipalities. To improve model fitting it was assumed $v[i] \sim t - Student(1, \xi, 2)$ where $\xi \sim Gamma(0.5, 0.005)$ for each municipality during year 1997. Figure 1 shows a comparison of the two posterior predictive p-values, with the normal distribution ($p - value = 0$) and the $t - Student$ distribution ($p - value = 0.635$).

From the 15 municipalities in Sucre state during the study period $1990 - 2002$, 7 of them presented relative risks greater than 1 in most years. These areas are mostly agricultural areas with poor living conditions.

References

Banerjee, S., Carlin, B. P., and Gelfand, A. (2003) *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall / CRC.

Gelfand, A.E., Ghosh, S. K. (1998) Model choice: a minimum posterior predictive loss approach. *Biometrika*, 85, pp. 1-11.

Lawson, A. B., Browne, W. J., and Rodeiro, C. L. V. (2003) *Disease Mapping with* WinBUGS *and MLwiN*, Wiley, New York.

Le, N. D., Zidek, J. (2006). Statistical Analysis of Environmental Space-Time Processes. Springer.

Spielgelhalter, D. J., Best, N., Carlin, B. P., and Van der linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J Roy. Statist.Soc., Ser. B*,64, 583-639.

Prediction of cancer mortality risks in spatio-temporal disease mapping [1]

Goicoa, T.[1], Ugarte, M.D.[1], Militino, A.F.[1], Etxeberria, J.[1,2]
[1] Department of Statistics and O. R., Universidad Pública de Navarra
[2] CIBER in Epidemiology and Public Health

email:tomas.goicoa@unavarra.es

Abstract:
The main goal of spatio-temporal disease mapping is describing the evolution of geographical patterns of mortality or incidence risks (rates). This could give clues to epidemiologists and public health researchers to formulate etiologic hypothesis of the disease. However, the ability of disease mapping models to make predictions about future mortality or incidence risks has not been widely explored. In this work, a flexible spatio-temporal model is considered for risk estimation and forecasting. The prediction MSE of both fitted and forecast values, as well as estimators of those quantities, will be derived. Spanish cancer mortality data will be used for illustration.

Keywords: P-spline models, CAR models, smoothing risks, forecasting.

1 Introduction

Health agencies plan cancer prevention resources based on cancer mortality/incidence risk estimations available to date. However, these official numbers are available after three or four years. In this context, statistical procedures providing mortality/incidence risk predictions for different regions or health areas are very useful. Using jointpoint regression models, Malvezzi et al. (2011) present estimates of mortality for all cancers and for selected major cancer sites in the year 2011 in the whole European Union and in its six more populated countries. They use actual mortality data up to the most recent available year, which is between 2005 and 2007 for most EU countries.

In this work flexible spatio-temporal models are considered to predict risks. The prediction MSE of both fitted and forecast values, as well as estimators of those quantities, will be derived. P-splines have been proposed in small areas to forecast dwelling prices (see Ugarte et al., 2009), and here, we extend this work to disease mapping spatio-temporal models including interaction terms. The methodology will

[1]This research has been supported by the Spanish Ministry of Science and Innovation (MTM 2008-03085/MTM).

be used to analyze several mortality cancer data for all the Spanish provinces in the period 1975-2008. Risks predictions for future years will be also provided.

2 Materials and Methods

Two models are considered for forecasting. Firstly, a model with CAR distributions for both the spatial and temporal random effects is considered. This model includes a spatio-temporal interaction term similar to those described by Knorr-Held (2000). Secondly, a spatio-temporal P-spline model described in Ugarte et al. (2010) is used. Smoothing is carried out in three dimensions: longitude, latitude, and time, allowing for different smoothing parameter in each dimension. Predictions will be obtained by extending the marginal time B-spline basis.

Consider n contiguous regions labelled $i = 1, \ldots, n$, and T time periods denoted by $t = 1, \ldots, T$. Conditional on the random region effects r_{it}, the number of deaths in each area and time period, C_{it}, is assumed to be Poisson distributed with mean $\mu_{it} = e_{it} r_{it}$, where r_{it} represents the unknown relative risks of mortality from a rare disease, and e_{it} is the expected number of deaths. Namely

$$C_{it} | r_{it} \sim Poisson(\mu_{it} = e_{it} r_{it}), \quad \log \mu_{it} = \log e_{it} + \log r_{it}. \tag{1}$$

In the spatio-temporal CAR model, the log-risk is modeled as

$$u_{it} = \log r_{it} = \beta + \phi_i + \gamma_t + \delta_{it}, \tag{2}$$

where β is an overall risk level, ϕ_i represents spatial effects, γ_t denotes temporal effects, and δ_{it} are space-time interaction effects. The distributions for the random effects $\boldsymbol{\phi}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$ are

$$\begin{aligned} \boldsymbol{\phi} &\sim N(\mathbf{0}, \sigma_s^2 \mathbf{D}_s) \quad ; \quad \mathbf{D}_s = (\lambda_s \mathbf{Q}_s + (1 - \lambda_s)\mathbf{I}_s)^-, \\ \boldsymbol{\gamma} &\sim N(\mathbf{0}, \sigma_t^2 \mathbf{D}_t) \quad ; \quad \mathbf{D}_t = \mathbf{Q}_t^-, \\ \boldsymbol{\delta} &\sim N(\mathbf{0}, \sigma_{st}^2 \mathbf{D}_{st}) \quad ; \quad \mathbf{D}_{st} = \mathbf{Q}_t^- \otimes \mathbf{Q}_s^-, \end{aligned}$$

where $\mathbf{Q}_s$ is determined by the spatial neighbourhood structure with the ith diagonal element equal to the number of neighbours of the ith region and for $i \neq j$, $\mathbf{Q}_{ij} = -1$ if i and j are neighbours and 0 otherwise; $\mathbf{I}_s$ is the $n \times n$ spatial identity matrix, and $\mathbf{Q}_t$ is determined by the temporal neighbourhood structure and it is analogously defined as $\mathbf{Q}_s$.

Model 2 can be expressed in matrix form as

$$\mathbf{u} = \mathbf{X}\beta + \mathbf{Z}_1 \boldsymbol{\phi} + \mathbf{Z}_2 \boldsymbol{\gamma} + \mathbf{Z}_3 \boldsymbol{\delta} = \mathbf{X}\beta + \mathbf{Z}\boldsymbol{\alpha}, \quad \boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{G}),$$

Using a P-spline spatio-temporal model the log-risk is modeled as

$$u_{it} = \log r_{it} = f(x_{1i}, x_{2i}, x_t), \tag{3}$$

where x_{1i} and x_{2i} are the coordinates of the centroid of the ith small area (longitude and latitude respectively), x_t is the time, and f is a smooth function to be estimated using P-splines with B-spline bases. One of the most interesting aspects of the P-spline models is that they can be expressed as linear mixed models using a one-to-one (orthogonal) transformation. Hence, the P-spline model can be represented as

$$\mathbf{u} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}, \quad \boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{F}^{-1}).$$

The $\mathbf{X}$ and $\mathbf{Z}$ matrices are given by

$$\begin{aligned} \mathbf{X} &= \mathbf{X}_3 \otimes (\mathbf{X}_2 \square \mathbf{X}_1) \\ \mathbf{Z} &= [\mathbf{Z}_1^* : \mathbf{Z}_2^* : \mathbf{Z}_3^* : \mathbf{Z}_4^* : \mathbf{Z}_5^* : \mathbf{Z}_6^* : \mathbf{Z}_7^*], \end{aligned}$$

and

$$\begin{aligned} \mathbf{Z}_1^* &= \mathbf{Z}_3 \otimes (\mathbf{X}_2 \square \mathbf{X}_1), \quad \mathbf{Z}_2^* = \mathbf{X}_3 \otimes (\mathbf{Z}_2 \square \mathbf{X}_1), \quad \mathbf{Z}_3^* = \mathbf{X}_3 \otimes (\mathbf{X}_2 \square \mathbf{Z}_1), \\ \mathbf{Z}_4^* &= \mathbf{Z}_3 \otimes (\mathbf{Z}_2 \square \mathbf{X}_1), \quad \mathbf{Z}_5^* = \mathbf{Z}_3 \otimes (\mathbf{X}_2 \square \mathbf{Z}_1), \quad \mathbf{Z}_6^* = \mathbf{X}_3 \otimes (\mathbf{Z}_2 \square \mathbf{Z}_1), \\ \mathbf{Z}_7^* &= \mathbf{Z}_3 \otimes (\mathbf{Z}_2 \square \mathbf{Z}_1), \end{aligned}$$

where the symbol $\square$ denotes the "row-wise" Kronecker product of two matrices (see for example, Eilers et al., 2006). Here, $\mathbf{X}_1 = [1 : \mathbf{x}_1]$, $\mathbf{X}_2 = [1 : \mathbf{x}_2]$, $\mathbf{X}_3 = [1 : \mathbf{x}_3]$, $\mathbf{Z}_1 = \mathbf{B}_1\mathbf{U}_{1s}$, $\mathbf{Z}_2 = \mathbf{B}_2\mathbf{U}_{2s}$, and $\mathbf{Z}_3 = \mathbf{B}_3\mathbf{U}_{3s}$. The matrices $\mathbf{B}_1$, $\mathbf{B}_2$, and $\mathbf{B}_3$ are the marginal B-spline bases for longitude, latitude and time; $\mathbf{U}_{1s}$, $\mathbf{U}_{2s}$ and $\mathbf{U}_{3s}$ come from the singular value decomposition of the penalty matrices for longitude, latitude and time, and the covariance matrix $\mathbf{F}^{-1}$ is a diagonal matrix arising from the representation of the P-spline model as a mixed model (see Ugarte et al., 2010 for more details).

The models are estimated using the well known penalized quasi-likelihood technique (PQL)(Breslow and Clayton, 1993). Risk predictions and their standard errors are obtained by extending the $\mathbf{X}$ and $\mathbf{Z}$ matrices.

3 Results

The methodology is illustrated analyzing prostate cancer in Spain from 1975 to 2008. Figure 1 displays relative risks estimates (1975-2008) and predictions (2009-2011) for four selected Spanish provinces, together with 95% confidence bands obtained with the P-spline model (3). A decreasing trend in mortality can be observed.

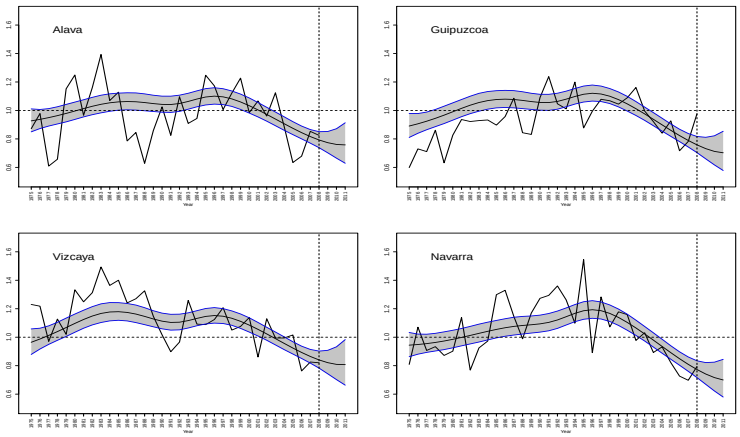


Figure 1: Smoothed prostate cancer mortality risks estimations and predictions with 95% confidence bands.

Acknowledgments

This research has been supported by the Spanish Ministry of Science and Innovation (MTM 2008-03085/MTM). The authors would like to thank to Marina Pollán from the National Epidemiology Center (area of Environmental Epidemiology and Cancer) for providing the data.

References

Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9-25.

Eilers, P.H.C., Currie, I.D., and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics and Data Analysis*, **50**, 61-76.

Knorr-Held L. 2000. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine* **19**, 2555-2567.

Ugarte M.D. , Goicoa T., Militino A.F., Durban M. (2009). Spline smoothing in small area trend estimation and forecasting, *Computational Statistics and Data Analysis*, **53**, 3616-3629.

Ugarte, M.D., Militino, A.F., and Goicoa, T. (2010). Spatio-temporal modelling of mortality risks using penalized splines. *Environmetrics*,**21**, 270-289.

Predictive assessment of a non-linear random effects model for space-time surveillance data

Michaela Paul, Leonhard Held
Department of Biostatistics, Institute of Social and Preventive Medicine,
University of Zurich

Abstract: Notification data collected by national surveillance systems are typically available as weekly time series of counts of confirmed new cases, stratified e.g. by geographic areas. This work outlines the statistical modeling framework in Paul and Held (2011) for the analysis of such data. Inherent (spatio-)temporal dependencies are incorporated via an observation-driven formulation. Using region-specific and possibly spatially correlated random effects, we are able to address heterogeneous incidence levels. Inference is based on penalized likelihood methodology for mixed models. The predictive performance of models is assessed using probabilistic one-step-ahead predictions and proper scoring rules.

Keywords: Time series of counts, infectious diseases, proper scoring rules.

1 Introduction

Notification data on infectious diseases typically consist of counts of confirmed new infections, which are observed in defined geographical areas at regular time intervals. Retrospective surveillance aims to identify outbreaks and (spatio-)temporal patterns through statistical modeling. Motivated by a branching process with immigration, Held et al. (2005) propose to decompose the mean incidence additively into three components: an *autoregressive*, a *neighbor-driven* and an *endemic* component. The first two components represent an autoregression on past counts in the same and in other regions, respectively, and should capture occasional outbreaks and dependencies across regions. The third component parametrically models regular trends and seasonal variation, e.g. by a sine-cosine formulation. Overdispersion can be allowed for by replacing the Poisson with a negative binomial distribution.

In the case of spatially correlated time series, the assumption of equal disease transmission or incidence levels across all regions is questionable. For instance, transmission might be influenced by age, vaccination status, or environmental conditions. Such factors could be incorporated into the model as covariates if suitable information is available. As an alternative, Paul and Held (2011) suggest to include regional random effects to allow for heterogeneity across regions. The predictive quality of the models is then investigated using one-step-ahead predictions and proper scoring rules (Gneiting and Raftery, 2007).

2 Methods

2.1 Modeling framework

Let y_{rt} denote the number of cases of a specific disease in region $r = 1, \ldots, R$ at time $t = 1, \ldots, T$. The counts are assumed to be Poisson or negative binomially distributed with conditional mean

$$\mu_{rt} = \lambda_r y_{r,t-1} + \phi_r \sum_{q \neq r} w_{qr} y_{q,t-1} + e_{rt}\nu_{rt}\,, \tag{1}$$

where $\lambda_r, \phi_r, \nu_{rt} > 0$ are unknown quantities, w_{qr} are suitably chosen known weights and e_{rt} corresponds to an offset (e.g. population numbers). A simple choice for the weights is $w_{qr} = 1$ if units q and r are adjacent and 0 otherwise.

The three unknown quantities are further decomposed additively on the log-scale and specified for example as

$$\log(\lambda_{rt}) = \alpha_0 + a_r \tag{2}$$

$$\log(\phi_{rt}) = \beta_0 + b_r \tag{3}$$

$$\log(\nu_{rt}) = \gamma_0 + c_r + \gamma_1 \sin(2\pi/52\, t) + \gamma_2 \cos(2\pi/52\, t) \tag{4}$$

where $\alpha_0, \beta_0, \gamma_0$ are intercepts, a_r, b_r, c_r are regional random effects, and the terms in curly brackets in (4) define the model seasonal variation. In applications, each of the three components may be suitably modified or omitted.

The stacked vector of all random effects is assumed to follow a normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. For instance, one may choose $\mathbf{\Sigma} = \mathbf{\Omega} \otimes \mathbf{I}$, where $\mathbf{\Omega}$ is an unknown 3×3 covariance matrix, and $\mathbf{I}$ is the $R \times R$ identity matrix. This formulation correlates the random effects (a_r, b_r, and c_r) between components, and leaves the random effects within each component (e.g., $\mathbf{c} = (c_1, \ldots, c_R)^\top$) uncorrelated.

In hierarchical models for spatio-temporal data, it is often reasonable to assume spatially correlated random effects rather than independent and identically distributed (iid) ones. Therefore, one might also adopt an intrinsic conditional autoregressive (ICAR) model (Besag et al., 1991) for the incidence levels $\mathbf{c}$, say. As the associated precision matrix has a rank deficiency of one, we apply a transformation $\mathbf{c} = \gamma_0 + \mathbf{Z}\tilde{\mathbf{c}}$ and estimate a reduced set of $R - 1$ random effects, $\tilde{\mathbf{c}}$, that are iid Gaussians (see Paul and Held, 2011).

The estimation of parameters involves integration of the likelihood with respect to the random effects which cannot be done analytically. Paul and Held (2011) suggest a penalized likelihood approach for inference, where variance components are treated as known when estimating the fixed and random effects. The variance components themselves are estimated through maximizing the approximated marginal likelihood obtained via a Laplace approximation.

2.2 Predictive model assessment

Model choice based on classical information criteria such as AIC is well explored and understood for models that correspond to fixed-effects likelihoods. However, their use can be problematic in the presence of random effects (Burnham and Anderson, 2002, p. 316). For model selection in time series models, the comparison of successive one-step-ahead predictions with the actually observed data is especially attractive. The often used mean squared error of several point predictions does not take prediction uncertainty into account. Instead, Gneiting and Raftery (2007) recommend the use of strictly proper scoring rules to evaluate probabilistic predictions in the form of a predictive distribution.

Strictly proper scoring rules simultaneously measure the sharpness and calibration of a prediction by assigning a numerical score based on a stated predictive distribution and the later observed actual value. The smaller the score, the better the predictive quality. Several proper scoring rules for count data are discussed by Czado et al. (2009). A popular scoring rule is the logarithmic score

$$\text{logS} = -\log(P(Y = y)) \tag{5}$$

which corresponds to the log predictive density at the observed value y. It is highly sensitive to extreme cases as it strongly penalizes low probability events. A more robust alternative is the ranked probability score

$$\text{RPS} = \sum_{k=0}^{\infty} \Big(P(Y \leq k) - 1(y \leq k) \Big)^2, \tag{6}$$

where 1 is the indicator function.

Typically, mean scores over a set of predictions are used to rank and compare different models informally or via tests such as a Monte Carlo permutation test for paired observations (see Paul and Held, 2011).

3 Case study

In a case study, Paul and Held (2001) applied the model to weekly influenza surveillance counts in 140 districts of Southern Germany for the years 2001–2008. Data were obtained from the SurvStat database of the Robert Koch Institute and analyzed using the functions implemented in the R package `surveillance` (Höhle, 2007). Exemplary R code to reproduce the analysis is given in the package vignette available at `https://r-forge.r-project.org/projects/surveillance/`.

The negative binomial model which yielded the lowest average logarithmic score, called 'B2', was specified by $\log(\lambda_{rt}) = \alpha_0$, $\log(\phi_{rt}) = \beta_0 + b_r$, and $\log(\nu_{rt}) = \gamma_0 + c_r + \gamma_1 t + \sum_{s=1}^{3} \gamma_{2s} \sin(2\pi s/52\, t) + \gamma_{2s+1} \cos(2\pi s/52\, t)$, where $(\mathbf{b}^\top, \mathbf{c}^\top)^\top \sim \text{N}(\mathbf{0}, \boldsymbol{\Omega} \otimes \mathbf{I})$ with $\boldsymbol{\Omega} = \begin{pmatrix} \sigma_b^2 & \rho\sigma_b\sigma_c \\ \rho\sigma_b\sigma_c & \sigma_c^2 \end{pmatrix}$. Here we consider a further model 'S', where the autoregressive

Model	$\overline{\text{logS}}$	$\overline{\text{RPS}}$
B2: with seasonal variation in (4)	0.5633	0.4363
S: with seasonal variation in (2) and (4)	0.5571	0.4224

Table 1: Average scores based on $140 \cdot 104$ one-step-ahead predictions.

component (2) additionally contains $S = 1$ seasonal terms. Average scores for this model, based on one-step-ahead predictions for years 2007–2008, can be found together with the scores for model B2 in Table 1.

4 Concluding remarks

The analysis showed that the predictive performance improves when the autoregressive parameter is also allowed to vary over time. In Paul and Held (2011), the inclusion of spatially correlated random incidence levels instead of iid ones did not substantially improve the predictive performance of a model which already incorporated spatio-temporal correlation via the neighbor-driven component.

References

Besag J., York J., Mollié A. (1991) Bayesian image restoration with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics*, 43, 1–20.

Burnham K. P., Anderson D. R. (2002) *Model Selection and Multimodel Inference. A Practical InformationTheoretic Approach*, Springer, New York.

Czado C., Gneiting T., Held L. (2009) Predictive model assessment for count data, *Biometrics*, 65, 1254–1261.

Gneiting T., Raftery A. E. (2007) Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association*, 102, 359–378.

Held L., Höhle M., Hofmann M. (2005) A statistical framework for the analysis of multivariate infectious disease surveillance counts, *Statistical Modelling*, 5, 187–199.

Höhle M. (2007) Surveillance: an R package for the monitoring of infectious diseases, *Computational Statistics*, 22, 571–582.

Paul M., Held L. (2011) Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts, *Statistics in Medicine*, 30, 1118–1136.

Selective Inference in Disease Mapping

Dolores Catelan, Annibale Biggeri
Dep. of Statistics "G. Parenti", University of Florence;
Biostatistic Unit, ISPO, Florence;
catelan@ds.unifi.it

Abstract: The main goal of Disease Mapping is to investigate the geographical distribution of the risk of diseases. Spatially-structured priors were considered in all the proposed models in the literature to estimate relative risk surfaces. Selective inference on area-specific relative risks received little attention in the literature. We refer to selection and estimation of relative risks of areas at unusual (higher and/or lower) risk. Previous use of cross-validation posterior predictive distributions to detect outlying observation misses to address the selection effect in inference. In this work we review this issue in the context of hierarchical Bayesian models and we take advantage of a real example on the distribution of Lung cancer in Tuscany.

Keywords: Cross-validation predictive distributions, hierarchical Bayesian model, Disease Mapping.

1 Introduction

Disease mapping, i.e. the study of variability of disease occurrence on space, focused on relative risk surface estimation. Since the seminal paper of Clayton and Kaldor (1987) spatially-structured priors were considered in almost all the proposed models in the literature. However, inference on area-specific relative risks received little attention in the literature despite of the need to select areas (or regions) at unusual (high or low) risk. Stern and Cressie (2000) used cross-validation posterior predictive distributions to explore model fitting and identify outlying areas in disease mapping. The idea of cross-validation is to re-fit the model removing one observation in turn. The model is thus fitted to a subset of data Y_{-i} from which the i-th observation is dropped. The posterior predictive distribution $P(Y_i^{rep}|Y_{-i})$ for a replicate (Y_i^{rep}) of the i-th observation conditional to the remaining data Y_{-i} is then used for evaluation purposes. The extremeness is usually measured by some summaries over $P(Y_i^{rep}|Y_{-i})$, for example the posterior predicted p-values, $P(Y_i^{rep} \leq y_i|Y_{-i})$, or the conditional predictive ordinate, $p(Y_i^{rep} = y_i|Y_{-i})$. Marshall and Spiegelhalter (2003) noted that "...There are essentially two reasons why observations/regions may be divergent. First, the statistical assumptions underlying the model may be incorrect...[second], these regions could represent genuine 'hot-spots' of disease requiring further investigation." Poor model fit is a reasonable explanation when a relevant number of observations/areas are identified as divergent while the presence of real hot-spots or outliers is the usual interpretation of few divergent ones.

Marshall and Spiegelhalter (2007) proposed a mixed approach to perform cross-validatory checks in disease mapping.

In this work we review this issue in the context of hierarchical Bayesian models and we take advantage of a real example on the distribution of Lung cancer in Tuscany.

2 Methods

Let Y_i be the number of observed cases in the i-th area ($i = 1, \ldots, 287$) which follows a Poisson distribution with mean $E_i\theta_i$, where E_i is the expected number of cases under indirect standardization and θ_i the relative risk.

Besag et al. (1991) specified a random effect log linear model for the relative risk $log(\theta_i) = u_i + v_i$. The heterogeneity random term u_i represents an unstructured spatial variability component assumed a priori distributed as Normal $(0, \lambda_u)$ where λ_u is the precision parameter modelled as Gamma. The clustering term v_i represents the structured spatial variability component assumed to follow a priori an intrinsic conditional autoregressive (ICAR) model. In other words, denoting S_i as the set of the areas adjacent to the i-th area, $v_i|v_{j\in S_i}$ is assumed distributed as Normal($\bar{v}_i$,$\lambda_v n_i$) where $\bar{v}_i$ is the mean of the terms of adjacent areas to the i-th one (Besag and Kooperberg, 1995) and $\lambda_v n_i$ is the precision, which is dependent on n_i, the cardinality of S_i. Through these two random terms the BYM model shrinks the relative risk estimates both toward the local and the general mean.

The choice of a suitable combination of hyperparameters leads to different degrees of prior vagueness on the extent relative risk heterogeneity among areas.

For the Besag et al. (1991) model we took advantage of the proposal of Bernardinelli et al. (1995). The hyperpriors for the precision parameters were parameterized in terms of the ratio between the 95th percentile and the 5th percentile of the relative risk distribution.

2.1 Cross-validation predicted p-values

Divergence from the hierarchical null models is assessed via posterior predictive distribution. The posterior predictive distribution is:

$$P(Y^{rep}|Y) = \int P(Y^{rep}|Y,\theta)P(\theta|Y)d\theta = \int P(Y^{rep}|\theta)P(\theta|Y)d\theta$$

assuming conditional independence of Y^{rep} and Y given the parameters. This is too confident since the data are used twice, for deriving posteriors and for obtaining replicates (Plummer 2008). To control for excess in optimism the posterior predictive distribution is replaced by the cross-validation (leave-one-out) posterior predictive distributions:

$$P(Y^{rep}|Y_{-i}) = \int P(Y^{rep}|\theta)P(\theta|Y_{-i})d\theta$$

Cross validation posterior predicted distributions are computationally prohibitive. Several approximations have been proposed. A mixed approach was given by Marshall and Spiegelhalter (2007). At each Montecarlo iteration a replicate value for the random parameters for the i-th observation is generated and then used to generate a replicate observation Y_i^{rep}. This approach is called mixed because random effects are drawn from their predictive distribution and not from the posterior.

A measure of divergence can be the cross validation posterior predicted p values defined, using mid-p for a discrete response, as:

- if $Y_i > E_i$: $Pr(Y_i^{rep} > Y_i^{obs}|Y_{-i}) + \frac{1}{2}Pr(Y_i^{rep} = Y_i^{obs}|Y_{-i})$
- if $Y_i < E_i$: $Pr(Y_i^{rep} < Y_i^{obs}|Y_{-i}) + \frac{1}{2}Pr(Y_i^{rep} = Y_i^{obs}|Y_{-i})$

where Y_i is the observed and E_i the expected number of cases in the i-th area.

The need of post-processing of any model-based p-values was discussed by Ohlssen et al. (2007).

3 Results

Lung cancer death certificates were considered for males resident in the 287 municipalities of the Tuscany Region (Italy) for the period 1995-1999. Data were made available by the Regional Mortality Register. A set of reference rates (Tuscany, 1971-1999) have been used to compute the expected number of cases for each municipality, following indirect standardization and classifying the population by 18 age classes (0-5, ..., 85 or more).

We explored several choices of hyperprior parameters for the Besag et al. model. These choices are expressed as prior 90% centile range of relative risk among areas. They represent different beliefs about the background variability of disease risk. Each choice produced a different nested set of divergent observations. The priors defined by the hyperparameters are very informative. In some sense, we deliberately specified a series of constrained bad-fitting models, which represents a series of believes on the role of confounders in modifying the baseline risk among areas. A vague (non informative) null with leave-one out (leave-a-group out) cross-validation did not work in our Disease mapping context.

4 Conclusion and Discussion

This approach does not correspond to a Bayesian version of hypothesis testing because a mixture model is not specified. One consequence is that posterior probabilities may not protect to multiple testing. Post-processing of cross-validation

posterior predictive p-values was used by Spiegelhalter. Tri-level Bayesian model was proposed by Catelan et al (2010) in the context of Disease Mapping. Similar approaches to hierarchical modelling of the null are described in Ohlssen et al. (2007). The authors argued that fitting null model by leave-one out cross-validation may be sufficient to detect divergent observations. We disagree with this point, as we show in the results section. In Disease mapping hierarchical modelling of the null can be reached by specifying informative null priors. Prior predictive, posterior predictive and partial predictive distribution can be discussed also in this context.

References

Bernardinelli, L., Clayton, D. and Montomoli, C. (1995). Bayesian Estimates of disease maps : how important are priors? *Stat Med*, **14**, 2411–31.

Besag, J., York,J., and Mollié, A. (1991). Bayesian Image Restoration, with Two Applications in Spatial Statistics (with discussion). *Annals of the Institute of Statistical Mathematic*, **43**, 1-59.

Catelan, D., Lagazio, C., and Biggeri, A. (2010). A hierarchical Bayesian approach to multiple testing in disease mapping. *Biometrical Journal* , **52**, 784-97.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disea-se Mapping. *Biometrics*, **43**, 671-81.

Lunn, D.J., Thomas, A., Best, N,. and Spiegelhalter D. (2000):WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, **10**, 4, 325-37.

Marshall, C.A. and Spiegelhalter, D.J. (2003). Approximate cross-validatory predictive checks in disease mapping models. *Stat Med*, **22**, 1649–60.

Marshall, C.A. and Spiegelhalter, D.J (2007). Identifying outliers in Bayesian hierarchical models: a simulation-based approach.*Bayesian Analysis*, **2**, 409-44.

Ohlssen, D.I., Sharples, L.D. and Spiegelhalter, D.J. (2007). A Hierarchical Modelling Framework for Identifying Unusual Performance in Health Care Providers. *Journal of the Royal Statistical Society, Series A*,**170**, 865-90.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**, 523-39.

Stern, H.S. and Cressie, N. (2000). Posterior predictive model checks for disease mapping models *Stat Med*, **19**, 2377–97.

A seismic swarm as a dynamic ergodic stochastic process: a case study of the L'Aquila's earthquake in 2009

Mauro Coli
Università degli Studi G. d'Annunzio, Pescara – Chieti-DMQTE
coli@unich.it

Abstract: The lethal earthquake of 6 April 2009 in L'Aquila, Central Italy, re-opened the discussion about the earthquake prediction due to the several precursory phenomena described in association to the event. One of the most important precursors that preceded L'Aquila main-shock was the foreshock activity. Papadopoulos et al. (NHESS, 2010) reported that a foreshock activity was there in the last months before the main-shock but the foreshock signal became very strong in the last 10 days with drastic changes in space-time-size domains of local seismicity. The importance of short-term foreshocks for the prediction of the main-shock was noted since the 1960's. However, foreshocks appear to precede only some main shocks and not others, while there are also foreshocks too small to detect by routine seismic analysis. In this context, the aim of the paper is to analyse the phenomenon of swarm as a dynamic ergodic stochastic process with particular reference to mean time of transition of a certain class of earthquake swarms (belonging to a certain state) to other classes of varying intensity. This kind of analysis can be referred to some indicators such as the mean first passage time and the mean time to return with their respective probabilities, that constitute an important interpretive tool in forecasting.

Keywords: seismic swarm, markovian processes, ergodicity.

1. Introduction

The available data consists of a data set of more than 15.000 shocks that occurred in the province of L'Aquila during the entire calendar year 2009 from 1st January to 31th December and were drawn by the Italian Seismic Bulletin (ISIDE).
A preliminary descriptive space time analysis of available data shows that the random phenomenon can be considered as a dynamic continuous parameter stochastic process and as such dealt with probability theory for the analysis of events random increments.
As known, the problem of ergodicity of a dynamic stochastic process has been addressed for the first time by physicists in the study of the kinetic theory of gases. For example, when a mass of gas is subject to random changes as a result of subsequent changes in status, the reiteration of these changes tends to create some regularity of behavior in the long run.
In our study of the swarm we will try to find out if there are similarities in its behavior to that of the kinetic theory of gases, using similar methods of analysis.

2. Methods and results

The dynamic characteristics of a destructive earthquake swarm, are known as being characterized by a foreshock (frequent shocks that occur before the main shock) and the main-shock, The shock of magnitude 6.3 that occurred on 6 th April can be placed within a sequence of four time intervals characterized as follow:

- Interval 1 - from January 1 to December 31, 2009 for a total of 15890 shocks;
- Interval 2 - from March 2 to May 2 (one month before and one month after the main-shock) for a total of 8611 shocks;
- Interval 3 - from March 22 to April 21, 2009 (fifteen days before and fifteen days after the main-shock) for a total of 6781 shocks
- Interval 4 - from 1 to 13 April 2009 (one week before and one week after the main-shock) for a total of 4369 shocks

For each of the four time intervals, we have defined five transition states corresponding to the following classes of earthquake magnitude:

State 1- (S_1)- shock with a magnitude of less than 1;
State 2 -(S_2) -shock with a magnitude between 1 and 1.4;
State 3 -(S_3) -shock with a magnitude between 1.4 and 1.8;
State 4 -(S_4) -shock with a magnitude between 1.8 and 2.4;
State 5-(S_5) -shock with a magnitude greater than 2.4.

From the above time intervals, we have estimated four 5x5 ergodic transition matrices and we have calculated the limit vectors and the corresponding matrix of the mean of first passage .
Dynamic processes are related to the time evolution and apply when the time factor (t) is a fundamental entity influencing the process.
In our study, states constitute a finite sequence of events not referred to the time at which they occurred.
An evolutionary system of random events is able to move between h incompatible transition states S_1, S_2, S_3, ..., S_i, ..., S_j, ..., S_h . At a given time the system may be in one and a only of these states. Once a certain state is reached at time (t_h), the system stays there until (t_k), with k steps of random transition , passes to the new state S_j .
In this case study, we are in the presence of a random evolutionary process and would like to know what is the probability that the system is in a generic state (S_i) with probability $p(S_i)$, regardless of the instant at which this happens, taking into account the type state previously occurred.
This can also be defined as the probability of transition from state S_i to S_j or P_{ij}. These probabilities are obtained studying the statistical behavior of the phenomenon: the frequencies with which state changes define an array whose elements correspond to the estimated transition probabilities, if normalized by row.
The “inheritance property” of few steps of transition, even if partial implying the system “memory”, may be limited.
For some classes of earthquake intensity the observed regularity allows to predict the future of the phenomenon and to conclude that some memory mechanism exists.

An effective way to verify the assumptions just mentioned, is to try to assess the situation after n successive steps of the transition process the ergodic behavior at the limit of its evolution.
Let P_{ij} denote the probability of transition from a single step, estimated with the observed data, the corresponding probability of transition P_{ij} (n) from i to j, in n steps. The transition may occur, in different ways, namely by following multiple mutually incompatible route A , B, or C, ...
The probability P_{ij} (n) is calculated as the sum of the probabilities of each route P_{ij} (n) = (p_{ij} (A)) + (p_{ij} (B)) + (p_{ij} (C)), where P_{ij} (n) gives rise a recurrence relation that consent us to distinguish some important features of the process during its evolution, such us the average transition time from one state to another or the average time to return to the starting state or even the time of permanence in a state, as well as the process configuration limit.
When the process is able to achieve any state of the system starting from any other during its evolution, it satisfies the conditions for ergodicity.
From an analysis of the indicators of the mean time of first passage for the four interval, we can see a substantial confirmation of the characteristics of the phenomenon in terms of probability of switching from one state to another.

Limit vector (15809 shocks)

S_1	S_2	S_3	S_4	S_5
0.01	0.67	0.13	0.10	0.07

Limit vector (8611 shocks)

S_1	S_2	S_3	S_4	S_5
0.01	0.14	0.34	0.30	0.21

Limit vector (6781 shocks)

S_1	S_2	S_3	S_4	S_5
0.01	0.14	0.34	0.30	0.21

Limit vector (4369 shocks)

S_1	S_2	S_3	S_4	S_5
0.046	0.082	0.295	0.344	0.274

Mean first passage time matrix (15809 shocks)

	S_1	S_2	S_3	S_4	S_5
S_1	13.16	4.48	3.94	6.072	9.22
S_2	13.12	4.18	3.78	6	9.21
S_3	13.19	4.31	3.67	5.83	9.19
S_4	13.2	4.5	3.8	5.93	9.06
S_5	13.2	4.52	3.92	6	9.03

Mean first passage time matrix (8611 shocks)

	S_1	S_2	S_3	S_4	S_5
S_1	156.24	1.24	8.54	13.87	24.48
S_2	169.27	1.49	8.01	12.88	23.46
S_3	175.42	1.76	7.63	11.59	22.86
S_4	176.25	2	9.01	9.9	20.16
S_5	176.84	2.17	10.32	11.09	14.28

Mean first passage time matrix (6781 shocks)

	S_1	S_2	S_3	S_4	S_5
S_1	12.19	3.32	3.2	6.63	24.48
S_2	12.64	3.7	3.05	6.44	23.46
S_3	12.75	4.22	3.03	5.91	22.86
S_4	12.82	4.74	3.65	5.05	20.16
S_5	12.94	5.08	4.29	5.83	14.28

Mean first passage time matrix (4369 shocks)

	S_1	S_2	S_3	S_4	S_5
S_1	217.38	10.95	3.67	4.71	1.15
S_2	227.02	12.19	3.27	4.72	1.04
S_3	227.84	13.7	3.33	4.4	0.9
S_4	228.25	14.88	4.13	3.98	0.18
S_5	228.72	15.78	4.9	4.67	0.08

For example, if we consider the 15.900 shocks, occurred in 2009, it would take 13 transitional stages to reach the transition state S_1 of lowest hazard from any previous

state, 5 steps to reach the state S_2, and so on, until 9 steps to be in the most dangerous state S_5 with a magnitude greater then 2.4.
A very different behaviour is observed for the seismic swarm of 4369 shocks occurred a week before and one after 6 April 2009.
During this time interval, shocks belonging to the state S_1 occurring very rarely and reaching the lowest value of the magnitude contemplated in the S_1 state took more than 200 stages of transition (shock). On the other hand, only one stage of transition is necessary in order to have two successively shocks of the highest magnitude S_5.
The number of stages of transition which determine the mean time to return from certain level to same level could be defined as an indicator dangerous due to recursion of this type of shock.

References

Imoto M. (1991) Changes in magnitude frequency b value prior to large ($M \geq 6$) earthquakes in Japan, *Tectonophysics*, 193, 311-325.

Papadopulos et al. (2010), Strong foreshocks signal preceeding the L'Aquila (Italy) earthquake (M_w 6.3) of 6 April 2009, *Nat. Hazards Earth Syst. Sci* , 10, 19-24, 2010.

Papadopulos et al. (2010) Identyfing seismicity levels via Poisson Hidden Markov Models, *Pure Appl.Geophis* , 167, 919-931.

Zambonelli et al. (2010) Performance of the Italian strong motion network during 2009. L'Aquila seismic sequence (Central Italy), *Bullettin of Earthquake Engineering*, 9, 39-65.

Geostatistical modeling of ice content within the "*Glacier Bonnard*" (Switzerland)

Nicolas Jeannée, Claire Faucheux[1]
GEOVARIANCES, 49bis av. Franklin Roosevelt, BP 91, 77212 Avon, FRANCE,
jeannee@geovariances.com

Eric Bardou, Pascal Ornstein[2]
CREALP rue de l'Industrie 45, 1950 Sion, SWITZERLAND

Abstract: The Bonnard glaciated mass (Valais region, Switzerland) overhangs the village of Zinal and its slow downward constant creep constitutes an environmental hazard. Ice content data have been acquired to better assess globally and locally the ice amount within the area, in order to evaluate the glacier's global dynamic and future evolution. Two ice content modeling approaches are tested: (i) a direct modeling using 3D simulations and (ii) to account for the relationship between the presence of ice and the lithology, a nested approach which consists in (1) simulating lithology with the plurigaussian method and then (2) populating each facies with ice content values. Both approaches are compared in terms of ice content prediction and of global ice mass.

Keywords: environmental hazard, Bonnard, ice content, facies modeling, plurigaussian.

1. Introduction

The "*Glacier Bonnard*" is a complex paraglacial complex located in the Canton of Valais (Switzerland). The glacier overhangs a settlement and its slow downward constant creep constitutes an environmental hazard. It is therefore important to understand the glacier's internal structure, particularly in terms of ice content, in order to evaluate its current global dynamic and future evolution.
Following preliminary geological and geophysical investigations, ice content data have been measured within several drillholes. Such data should allow to assess globally and locally the amount of ice within the "*Glacier Bonnard*" area.
Two geostatistical modeling approaches are considered: a direct modeling of ice content and an indirect approach which accounts for the relationship between lithology and ice content. Both approaches are presented and compared in terms of ice content prediction and of global ice mass among the sampled area.

2. Material

The studied area is part of catchment that ranges between the altitude of 2750 and 3000 m (cf. Fig 1). At this geographical location, cold temperature (freezing) and snow play an important role in the annual water balance. Approximately 80 % of the surface is composed by creeping permafrost. The geology of the source area is located on the contact between the thrust sheets of the Dent Blanche and Tsaté systems (Pilloud & Sartori, 1981). The granitic gneisses of the Dent Blanche covering the glacier Bonnard fall from the cliffs that surround summital crests. The outcrops and the cliffs are much fissured and produce a very large amount of blocks.
As basis for determining the bedrock top, 11 lines of refraction seismic have been acquired and treated in tomography due to the chaotic relief. Even if the overall results

seem to be consistent, small discrepancies were found between the lines (artefacts, unlikely geological features). A 3D managing sources software (Adhoc 3D solutions) has then been used to fit seismic refraction tomographic sections with drillholes and field geological mapping information. Along the refraction lines, points representing the bedrock top have been generated every 5 m. Those points are used to make an ordinary kriging of the bedrock top.

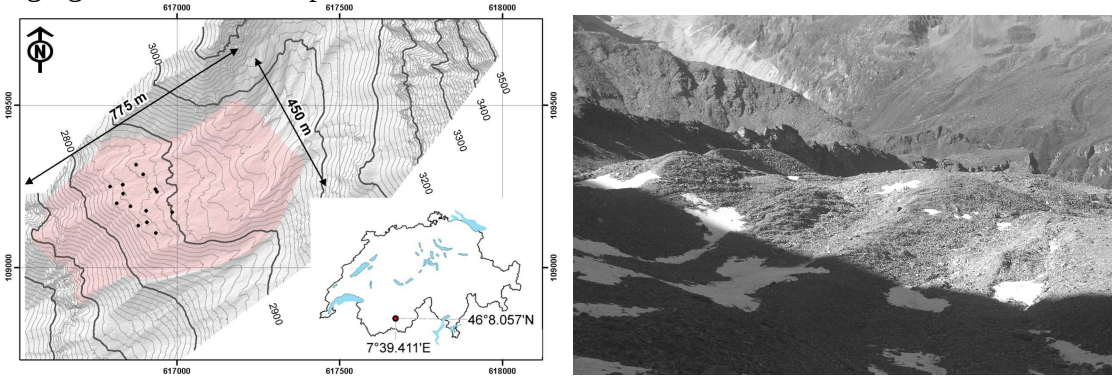


Figure 1. Left: global view of the studied area with drillholes location (black dots). Right: view from the top of the studied area in mid-summer 2009.

Fourteen boreholes have then been drilled using the Down-The-Hole Drill method (Fig. 1). This is a fast but completely destructive way to drill, which makes interpretation quite difficult. Indeed the only material available is cuttings smaller than 2 cm. All this information helped in determining homogeneous areas in terms of glaciated mass behavior, as illustrated on Fig. 2.

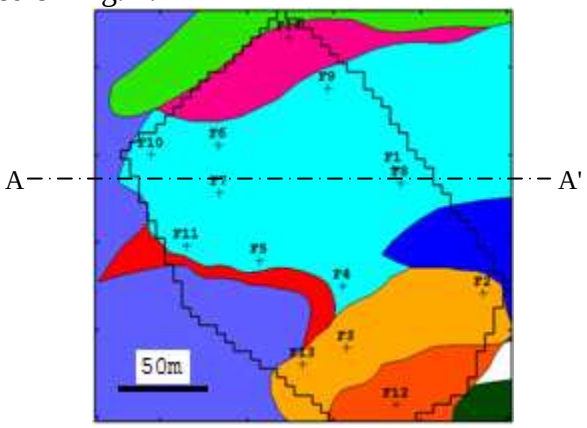


Figure 2. Outline of the studied area (broken line), drillholes (+) and polygons displaying homogeneous areas regarding the glaciated mass behavior.

3. Methodology

Several approaches might be applied to assess locally the ice content distribution and its variability. 3D conditional simulations have been first performed using the Turning Bands approach, after a Gaussian anamorphosis transformation (Chilès & Delfiner, 1999). This approach accounts for the spatial variability of the ice content, captured by a classical variogram analysis.

A strong relationship is expected between the lithotype and the ice content. Therefore, in order to account for that, 3D facies simulations are computed using the Plurigaussian algorithm (Armstrong et al., 2003), which allows integrating the geological knowledge about the expected transitions between facies (inferred from field survey). The algorithm consists first in determining the proportions of each lithothype over the 3D domain using drillholes information. The proportions are locally modified using the

definition of homogeneous areas. Then a lithotype rule is chosen (Fig. 3) to describe the relationship between the facies. Variogram models for the two Gaussian variables are fitted such as to reproduce the spatial continuity of the lithotypes. Finally, Gaussian variables are simulated and truncated so as to get the facies simulations (Fig. 3).

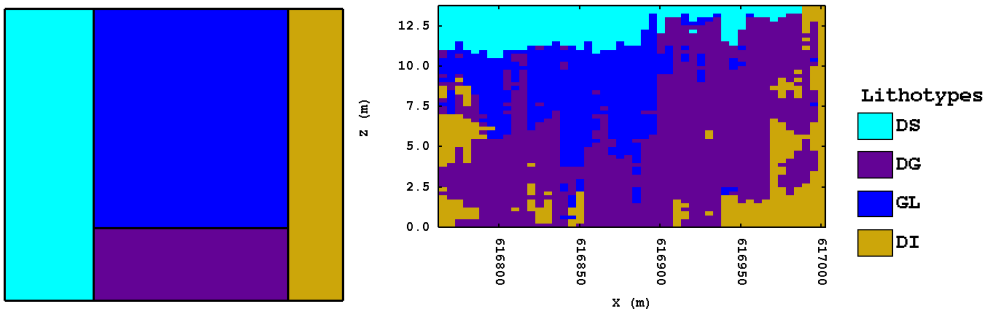


Figure 3. Left: Lithotype rule displaying authorized transitions between the facies. Right: Example of a lithotype simulation (along A-A' on Fig. 2) using plurigaussian (see Table 1 for lithotype meaning).

Once obtained, the facies simulations are populated with ice content values. The distribution of ice content within each facies is assumed random and to follow a triangular distribution.
For both approaches, a preliminary flattening has been performed to increase the lateral consistency of facies and ice content data. Once the 3D ice content simulations are obtained in the flattened space and converted back to the structural space, a post-processing is applied to determine the global distribution of ice content mass. The results will interestingly be compared with the computation of a statistical global mean and standard deviation, which assumes that the ice content is purely random within the area of interest.

3. Results & Discussion

The global simulated volume, for the area of interest (Fig. 2), is equal to 465 000 m^3. Classical statistics show the strong link between lithotypes and ice content (Table 1). Merging all data together contributes to largely increase the ice content variability.

Lithotype	Count	Min.	Q50	Max.	Mean	Std. Dev.
All	352	0	10	95	29.02	34.37
Superficial diamict (DS)	48	0	0	0	0.00	0.00
Glaciated diamict (DG)	137	0	20	60	21.42	13.18
Ice (GL)	86	60	87.5	95	84.65	10.45
Diamict (DI)	81	0	0	0	0.00	0.00

Table 1. Elementary statistics of ice content (in %) globally and for each lithotype.

Fig. 4 shows one ice content simulation obtained for each approach (turning bands simulation and plurigaussian simulation followed by a population with ice content values). Differences clearly appear: with the plurigaussian simulations ice content patterns are well defined due to the consideration of lithology. Furthermore, the horizontal changes suggest the presence of 2-3 ice bodies of which at least 2 are almost disjoint. Those features are consistent with field observations.
Finally, global estimates of ice mass are displayed in Table 2 within the area of interest. Similarity between the means is obvious whereas the standard deviations are very different. The statistical approach ignores data redundancy due to the spatial continuity

and therefore underestimates the variability. Regarding geostatistical simulations, TB overestimates standard deviations; indeed, lithology being ignored, a lot of unrealistic intermediate values are simulated. On the contrary, PGS integrates the lithology and therefore produce more realistic results. PGS fits better with the field knowledge, particularly in producing simulations presenting discrete elements (ice-bodies, lateral moraine, etc.). These results are coherent with the genesis of those different features that are more colliding than mixing together.

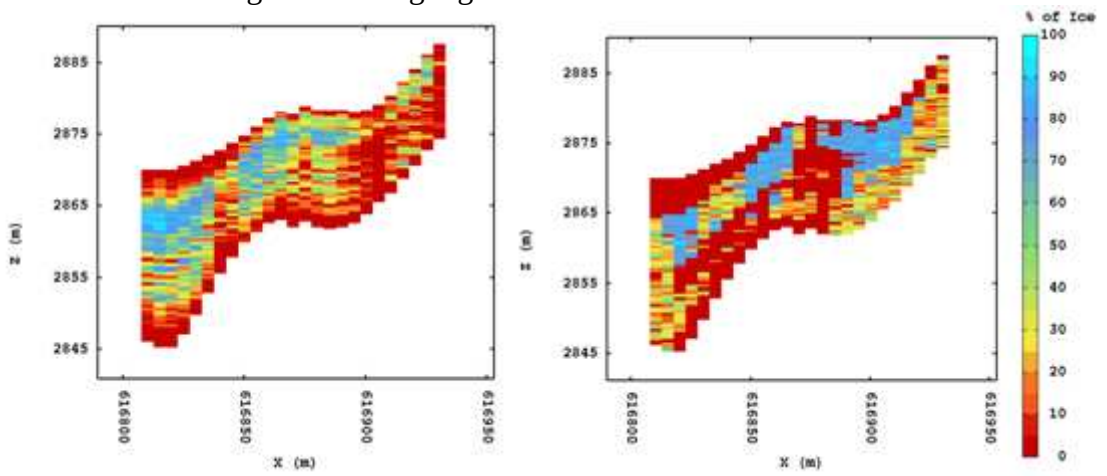


Figure 4. Cross-section along A-A' of the ice content simulations using turning bands (left) or indirectly via plurigaussian (right),for the same simulation.

Approach	Mean (kt)	St. Dev. (kt)	CV (%)	Q5 (kt)	Q95 (kt)
Statistics	123.69	2.26	1.83%		
Direct (TB)	130.08	21.87	16.81%	90.45	166.22
Indirect (PGS)	129.61	10.82	8.35%	111.08	147.33

Table 2. Classical statistics related to the global ice content mass estimated within the area of interest (in kilo tons of ice).

4. Concluding remarks

In this context of complex material with quick spatial changes, classical approaches like the single use of geophysics failed in providing an appropriate framework for detailed hazard assessment. Plurigaussian simulations allowed quantifying the total ice mass while taking into account the available information (lithology, homogeneous areas).

Acknowledgments

The authors wish to thank the State of Valais for supporting data acquisition, our colleague Guillaume Favre-Bulle for field work and the project INTERREG IV RiskNat (2009-2013) to provide funding for this study.

References

Armstrong M., Galli A., Le Loc'h G., Geffroy F and Eschard R. (2003), *Plurigaussian Simulations in Geosciences*, Springer, 160 p.

Chilès J.P. and Delfiner P., (1999), *Geostatistics: Modeling spatial Uncertainty*, John Wiley & Sons, 720 p.

Pilloud, C., Sartori, M. (1981), *Etude géologique et pétrographique de la région des Diablons (Val de Zinal, VS)*. MSc. Thesis, Université de Lausanne.

Is space-time interaction real or apparent in seismic activity?

Renata Rotondi, Elisa Varini
C.N.R. - Istituto di Matematica Applicata e Teconologie Informatiche, Milano (I),
reni@mi.imati.cnr.it

Abstract: It is widely shared opinion that not only secondary (aftershocks) but also main earthquakes have the tendency to occur in space-time clusters. This assumption has affected the preferential choice of stochastic models in the studies on seismic hazard, like self-exciting (epidemic) models which imply the abrupt increase of the occurrence probability after a shock and the subsequent exponential decrease without the desirable increase before a forthcoming event. The importance of this assumption requires the application of statistical tools to evaluate objectively its coherence with the reality at different scale of magnitude-space-time. To this end we consider the earthquakes drawn from the historical Italian catalogue CPTI04 that geologists have associated with each of the eight tectonically homogeneous regions in which Italian territory is divided. Fixing different magnitude thresholds we perform statistical tests based on the space-time distance between pairs of earthquakes under the null hypothesis of uniform distribution in time and space and evaluate the significance of the possible clusters. Monte Carlo hypothesis testing is also used to obtain the null distribution and the simulated p-value.

Keywords: detection of space-time clusters, Knox test, K-nearest neighbour test, Mantel test

1 Introduction

Some occurrence patterns in the worldwise seismicity are ascribable to space-time clustering; the best-known is due to the aftershocks, smaller earthquakes that follow a previous large shock within a distance up to twice the rupture length from the mainshock and can continue over a period of weeks, months, or years. Some articles in the literature claim that also strong events occur in clusters (Kagan and Jackson (2000), Lombardi and Marzocchi (2007)); this feature, if validated, would have heavy consequences on the choice of models in hazard assessment. We think that it is necessary to pass from quantitative observations to inferential tests which assign the statistical significance to some assumptions. Three types of tests can be carried out (Rogerson and Yamada (2009)): general and focused tests and tests for the detection of clustering. General tests provide a global statistic that assesses the degree to which a pattern deviates from the null hypothesis of space-time randomness without giving informaton on the size and location of clusters, focused tests are used to know

whether a cluster exists around prespecified foci, whereas in the third category many local tests are carried out simultaneously to uncover the location and size of any possible clusters by scan-type statistics. This article concerns the first step of a study on Italian seismicity in which we try to answer the question whether, for given magnitude thresholds, the global pattern of the past seismicity in tectonically homogeneous Italian regions is significantly clustered. In the future, where the answer is positive, we are going to establish, by scan-type statistics, whether the study region is homogeneous, and, where the answer is negative, to uncover isolated hot spots of increased activity and to look for geophysical explanations of this fact.

2 Space-time tests on tectonic regions in Italy

We consider three global tests: Knox, Mantel and Jacquez (or k NN) tests (Tango (2010)). Knox's statistic counts the number of observed pairs of n events close in both space and time:

$$T = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^S a_{ij}^T \tag{1}$$

where

$$a_{ij}^S = \begin{cases} 1, & i \neq j \text{ and } d_{ij}^S < \delta_1 \text{ (km)} \\ 0, & \text{otherwise} \end{cases} \qquad a_{ij}^T = \begin{cases} 1, & i \neq j \text{ and } d_{ij}^T < \delta_2 \text{ (years)} \\ 0, & \text{otherwise} \end{cases}$$

and δ_1 and δ_2 are unknown critical space and time limits to be prespecified. Under the null hypothesis H_0 - the temporal distances between pairs of events are independent of the spatial distances - it is proved that mean and variance of T are given by:

$$E(T) = \frac{N_{1S}\ N_{1T}}{N}$$

$$Var(T) = \frac{N_{1S}\ N_{1T}}{N} + \frac{4\ N_{2S}\ N_{2T}}{n(n-1)(n-2)} - \left(\frac{N_{1S}\ N_{1T}}{N}\right)^2 + \frac{\{N_{1S}(N_{1S}-1) - N_{2S}\} \times \{N_{1T}(N_{1T}-1) - N_{2T}\}}{n(n-1)(n-2)(n-3)}$$

where $N = n(n-1)/2$, $N_{1S} = \frac{1}{2}\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^S$ and $N_{2S} = \frac{1}{2}\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k\neq j}^{n} a_{ij}^S\ a_{ik}^S$ (analogously we get N_{1T} and N_{2T} substituting a^S with a^T). Given values of δ_1, δ_2 and observed $T = t$, the null distribution of T and its p-value can be approximated by either one of the following:

• Poisson distribution when N_{1S} and N_{1T} are small compared with N (or $E(T)$ is roughly equal to $Var(T)$) with

$$\text{mid}-p-\text{value} = 1 - \sum_{k=0}^{t} \frac{E(T)^k}{k!}\ exp\{-E(T)\} + \frac{1}{2}\ \frac{E(T)^t}{t!}\ exp\{-E(T)\}$$

• Normal distribution with p-value given by: $1 - \Phi\left(\frac{t-E(T)}{\sqrt{var(T)}}\right)$

• Monte Carlo hypothesis testing: we simulate the null distribution of T calculating the same statistic for a large number N_{rep} of data sets obtained by permuting the times among the fixed spatial locations (or viceversa). In this way we get:

$$\text{Simulated } p\text{−value} = \frac{1 \ + \ \sum_{\nu=1}^{N_{rep}} I(T_\nu \geq T_{obs})}{N_{rep} + 1} \qquad . \tag{2}$$

Mantel's test is a generalization of the Knox's test based on the same statistic (1) where reciprocal transformations of the distances are used to increase the influence of close distances and decrease that of the long distances, hence we have:

$$a_{ij}^S = \frac{1}{d_{ij}^S \ + \ c_1} \quad (a_{ii}^S = 0) \qquad\qquad a_{ij}^T = \frac{1}{d_{ij}^T \ + \ c_2} \quad (a_{ii}^T = 0)$$

with c_1 and c_2 unknown constants. To avoid the issues concerning the choice of the δ and c constants, Jacquez proposed a Knox-type test where the closeness is defined by the k nearest neighbours (k NN) such that:

$$a_{ij}^S = \begin{cases} 1, & \text{if event } j \text{ is a } k \text{ NN of event } i \ (\neq j) \text{ in space} \\ 0, & \text{otherwise} \end{cases}$$

Analogously we get a_{ij}^T. Monte Carlo hypothesis testing is required to obtain the null distribution of T and the simulated p-value (2) for both the Mantel's and the Jacquez's test.

3 Results

We have applied these tests to two data sets constituted by the 383 and 45 earthquakes of magnitude $M_w \geq 4.5$ and $M_w \geq 5.3$ respectively, occurred in the Central Northern Apennines West region characterized by normal faults. Figures 1 and 2 synthesize graphically some results of Knox's and Jacquez's tests showing the p-values obtained as the constants of the tests vary. We point out that space-time clustering of earthquakes of $M_w \geq 4.5$ is statistically significant for some values of δ_1, δ_2 and k, but it isn't when the threshold increases; consistent results are also obtained through the Mantel's test. This means that in the Italian tectonic context space-time clustering is not a property invariant to the magnitude threshold contrary to what is stated in the literature (Lombardi and Marzocchi (2007)). Hence this property must be verified through statistical tests so that the most appropriate stochastic model for hazard evaluation is proposed in each specific context.

References

Kagan Y.Y. and Jackson D.D. (2000) Probabilistic forecasting of earthquakes, *Geophysical Journal International*, 143, 438-453.

Lombardi A. and Marzocchi W. (2007) Evidence of clustering and stationarity in the time distribution of large worldwide earthquakes, *Journal Geophysical Research*, 112, B02303, doi:10.1029/2006JB004568

Rogerson P. and Yamada I. (2009) *Statistical Detection and Surveillance of Geographic Clusters*, Chapman & Hall.

Tango T. (2010) *Statistical Methods for Disease Clustering*, Springer, New York.

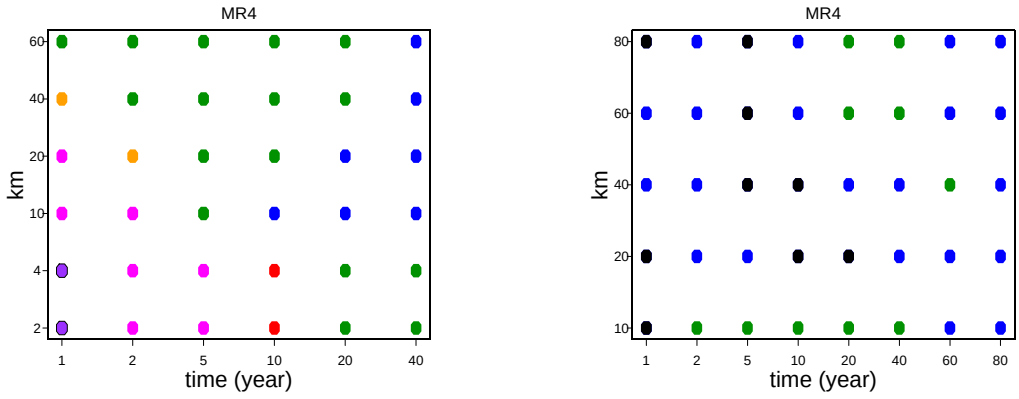


Figure 1: p-value of the Knox's test applied to earthquakes of $M_w \geq 4.5$ (left) and $M_w \geq 5.3$ (right) for different values of δ_2 (x-axis) and of δ_1 (y-axis): $p \leq 10^{-6}$ (violet), $10^{-6} < p \leq 0.01$ (magenta), $0.01 < p \leq 0.05$ (red), $0.05 < p \leq 0.10$ (orange), $0.10 < p \leq 0.50$ (green), $0.50 < p \leq 0.95$ (blue), $p > 0.95$ (black).

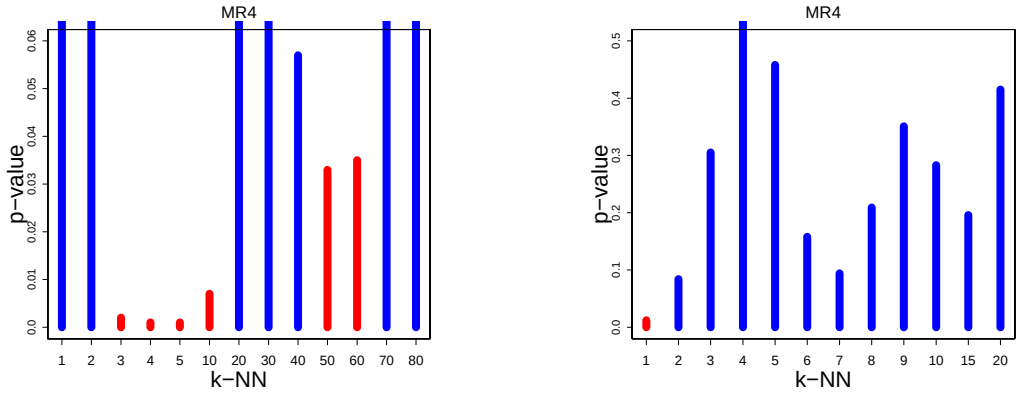


Figure 2: p-value of the Jacquez's test applied to earthquakes of $M_w \geq 4.5$ (left) and $M_w \geq 5.3$ (right): $p \leq 0.05$ (red), $p > 0.05$ (blue). Order k of the nearest neighbours on the x-axis.

Spatio-temporal modelling for avalanche risk assessment in the North of Italy

Orietta Nicolis
University of Bergamo
Dept. of Information Technology and Mathematical Methods
E-mail: orietta.nicolis@unibg.it

Assunção, Renato
Universidade Federal de Minas Gerais, Departmento de Estatística
Belo Horizonte, MG, Brazil
E-mail: assuncao@est.ufmg.br

Abstract: The main objective of this work is to evaluate the avalanche activity in a given location and at a given time taking into account a number of variables including the stratigraphy of snow cover, temperature, direction and wind speed, altititude, etc. To this end we propose a space-time point model where the intensity function indicates the limiting expected rate of occurrence of snow avalanches of a given size occurring on a certain day at location (x, y), conditioned on the historical information available prior to time t. Some meteorological and environmental data may be considered as the covariates of the model. To show the ability of the model in assessing the risk avalanche, data from digitalized Avalanche Database of the Trentino Region (North of Italy) is considered. Since not all locations in the Alpine zone are equally likely subject to snow avalanche, the model will be flexible enough for including a spatially-varying background rate of avalanche which may be estimate by kernel smoothing the observed avalanches .

Keywords: Snow avalanches, intensity function, spatio-temporal modelling.

1 Introduction

In recent years the study of avalanche phenomena has attracted growing interest especially for the increase of accidents and deaths, now comparable with those related to natural disasters. This is mainly due to a wide anthropization of mountain areas which has often brought a rapid growth of recreational activities, transportation, and constructions in high-altitude areas without an adequate assessment of avalanche hazard. Hence, the analysis of avalanche activity is extremely important to prevent damage and for activities aimed at land use planning in mountain areas. Many scientists have been studying avalanches to try to map the risk and improve predictions. To that end several statistical methods have been proposed based on different approaches. In this work we propose an approach based on space-time

point processes for modeling the avalanche risk. In particular, the intensity function of the process indicates the limiting expected rate of occurrence of snow avalanches occurring on day t at location (x, y), conditioned on the historical information available prior to time t. Also, we use a self-exciting model to deal with unobserved random space-time effects. The location (x, y) represents the baricenter of the polygon which draws the shape of avalanche. For showing the effect of some covariates (such as elevation, slope, temperature, etc.) different models are proposed. Application to the digitalized Avalanche Dataset of Trentino region (Italy) illustrates the ability of the models to forecast the risk avalanche. Although this approach has not been previously applied to avalanche events, it has been used for analysis spatio-temporal analysis of earthquakes occurrences (Ogata, 1998) and wildfire risk (Peng *et al.* 2005; Schoenberg *et al.* 2007).

2 Spatio-temporal models for avalanches

Any spatial temporal point process is uniquely characterized by its conditional intensity function $\lambda(t, x, y|\mathcal{H}_t)$ given by the limiting conditional expectation

$$\lambda(x, y, t|\mathcal{H}_t) \lim_{\Delta t, \Delta x, \Delta y \downarrow 0} \frac{E[N\{(t, t+\Delta t) \times (x, x+\Delta x) \times (y, y+\Delta y)\}|\mathcal{H}_t]}{\Delta t, \Delta x, \Delta y}$$

provided the limit exists. This is a random function that depends on the prior history, $\mathcal{H}_t$, of the point process up to time t. In this preliminary analysis, we considered a small number of models that should capture the main aspects of the avalanche dataset. One first class of models is nonparametric and has separable spatial and temporal effects. This is given by

$$\lambda_{1a}(x, y, t|\mathcal{H}_t) = \lambda(x, y, t) = \beta_0 + \beta_1 S(x, y) + \beta_2 T(t) \tag{1}$$

or by

$$\lambda_{1m}(x, y, t|\mathcal{H}_t) = \exp\left(\beta_0 + \beta_1 S(x, y) + \beta_2 T(t)\right) \tag{2}$$

where β is the parameter vector to be estimated. So, one is an additive model while the other is a multiplicative model. In these models, $S(x, y)$ is a deterministic function of the location (x, y) and it is estimated by a two-dimensional kernel smoother

$$S(x, y) = \frac{1}{n_0} \sum_{j=1}^{n_0} K\left(\frac{x - x_{0j}}{\phi_x}\right) K\left(\frac{y - y_{0j}}{\phi_y}\right)$$

where K is a suitable kernel function, taken as the quartic kernel in this paper. The function $T(t)$ is a periodic with trend deterministic function, also estimated by kernel methods using the events' times. The determinist aspect of these functions make the conditional intensity independent of the past, justifying the first equality in (1). To have an identifiable model and to avoid numerical instabilities, we centered all covariates at zero. It is likely that this model has less predictive power than

other models as it does not incorporate important additional information. However the model can be improved using covariates. At this moment, we have the elevation $E(\mathbf{x})$ and slope $S(\mathbf{x})$. In particular, for the slope we created a binary map with areas with slopes angles within this (25, 50) degrees. Hence, another class of models has an intensity varying only with the exogenous covariates and the temporal components. We again have $\lambda(x, y, t|\mathcal{H}_t) = \lambda(x, y, t)$ for these models, a deterministic intensity function. It is given by

$$\lambda_{2a}(x, y, t) = \lambda_{1a}(x, y, t) + \beta_3 E(\mathbf{x}) + \beta_4 S(\mathbf{x}) \tag{3}$$

Another version of this model is the multiplicative form where

$$\lambda_{2m}(x, y, t) = \lambda_{1m}(x, y, t) \exp\left(\beta_3 E(\mathbf{x}) + \beta_4 S(\mathbf{x})\right) \tag{4}$$

Other covariates can be added in the model such as precipitation, temperature and the level of new snow. Additional improvements of these models respect to the first class of models can be tested by means of the difference between the log-likelihood maximum values of each model. The final class of models we are going to consider are those that include the history of previous avalanches events in the area near each point. The conditional intensity is a truly random function that depends on the previous occurrences. Let

$$H(\mathbf{x}, t) = \int\int\int I_{B_{\mathbf{x}}(r)\times[t-\epsilon,t)}(x, y, t) N(dx, dy, dt)$$

where $I_A(\cdot)$ is the indicator function of the set A and $B_{\mathbf{x}}(r)$ is a small disc centered at $\mathbf{x}$ and with radius r. That is, $H(\mathbf{x}, t)$ is the number of events from the point process N that are inside the three-dimensional cylinder $B_{\mathbf{x}}(r) \times [t - \epsilon, t)$. Clearly, $H(\mathbf{x}, t)$ is $\mathcal{H}_t$-measurable. Then, the models incorporating this previous history are of two types, an additive model,

$$\lambda_{3a}(x, y, t|\mathcal{H}_t) = \lambda_{2a}(x, y, t) + \beta_6 H(\mathbf{x}, t)\,, \tag{5}$$

and its multiplicative version,

$$\lambda_{3m}(x, y, t|\mathcal{H}_t) = \lambda_{2m}(x, y, t) \exp(\beta_6 H(\mathbf{x}, t))\,. \tag{6}$$

3 Applications and results

The data used in this work have been provided by the province of Trento through the availability of digitalized Avalanche Database (based on a permanent survey on avalanches). In this application we consider 3350 avalanche events at 970 sites for the period January 1980 – December 1989. In this preliminary report, we did not fit the models (5) and (6). They require a much heavier numerical work as each time unit (day, in our case) has an associated map with the covariate $H(\mathbf{x}, t)$ that

Models	Intercept	$S(\mathbf{x})$	$T(t)$	Elevation	Slope	Log-Lik
Model 1	0.16017	0.00027	0.01559	NA	NA	-1803.152
Model 2	0.05780	0.00027	0.15635	0.23963	0.00035	-1138.234

Table 1: Estimates from models 2 and 4.

enters the likelihood maximization in each iterative step. We are working on this model and should have final results soon. The results for the models 2 and 4 are in Table 1. Figs. 1 (middle and right) show an example of the estimated intensity functions (risk maps) by the two models on February 1, 1986. As expected, model

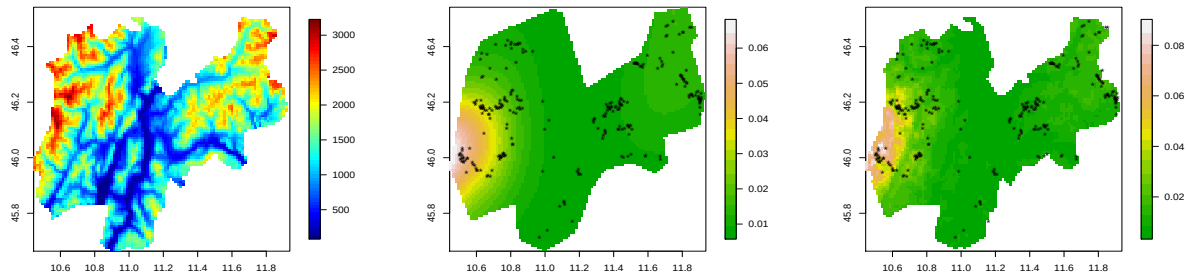


Figure 1: Elevation map of Trentino (left); estimated intensity function at February 1, 1986 by model (2) (middle) and model (4) (right). Asterisks represent avalanche events at the same day.

(4) performs better than model (2). We are going to include other covariates such as temperature and the amount of snow accumulated in the soil. Both are time varying and should be useful in terms of prediction of avalanche events. We are in the process of collecting these covariates and we expect to have an extended version of this paper incorporating these additional information in the near future.

References

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *An- nals of the Institute for Statistical Mathematics* (50), 379-402.

Peng, R. D., Schoenberg, F. P., Woods, J. (2005). A space-time conditional intensity model for evaluating a wildfire hazard index. *Journal of the American Statistical Association*, 100 (469), 26–35.

Schoenberg, F., Chang, C., Keeley, J., Pompa, J., Woods, J., and Xu, H. (2007). A critical assessment of the Burning Index in Los Angeles County, California. *International Journal of Wildland Fire*, 16, 473-483.

A data driven model for spatio-temporal estimation of shallow water table depth in soils

Fabrizio Ungaro, Costanza Calzolari

CNR IRPI Firenze, Via Madonna del Piano 10, 50019 Sesto F.no (FI),
fabrizio.ungaro@irpi.cnr.it

Abstract:
Data from a monitoring network were used to develop a data driven model for predicting water table depth in space and time. Records of 160 piezometer sites available from 1997 to 2010 were analyzed to detect the overall temporal trend in water table depth in a relevant agricultural area in Northern Italy. Evolutionary Polynomial Regressions (EPR) were used to calibrate a predictive tool based on climatic data and the records from 47 selected sites between 2004 and 2009 (N = 5611). The model was validated against the WT depths observed in 15 independent sites between 2005 and 2010 (N= 2052). Validation resulted in a mean absolute error of 30.0 cm (R^2 = 0.65). The general model was extended to the whole area, using the geostatistical estimates of the average water table depth as input, providing spatio-temporal maps of the water table depth at any give date.

Keywords: Soil water table; Evolutionary Polynomial Regressions; Spatio-temporal maps

1. Introduction

In alluvial plain areas of Northern Italy, the presence of seasonally saturated horizons due to the presence of a shallow (0.5- 3 m) groundwater within the rooting depth of crops, during the growing period, allows optimizing the irrigation water supply, thus saving the resource, concentrating the irrigations only where and when really needed. To this aim, operational tools are needed to estimate the depth of the shallow water table with a good reliability in space and time. Predictions about shallow water table dynamics can be made adopting different approaches (Morgan and Stolt, 2004). These can be either based on i) climatic records, ii) on field evidences from soil morphology and properties, iii) on the outputs of physically based water balance models or iv) on combinations of the different approaches. Water balance models require a high load of input data, which are not always available, and, especially in the case of spatially distributed models, the reliability for applicative goals is not always assured, being this often related to site specific soil conditions (Salazar et al, 2008). Empirical modelling can represent a suitable alternative, providing good results when locally calibrated. The choice among the available approaches is often determined by the density and frequency of observations over the space and time domains. Aim of this work is to obtain reliable estimates of water table (WT) depth both in space and time to be used for crop irrigation water requirement assessment, using a limited amount of climatic and hydrological information. An empirical time series model is presented, based on

meteorological data and water table depth records irregularly spaced in time and space. The model, calibrated in alluvial plains of Emilia Romagna (Italy), is based on a limited number of piezometer wells and climatic data and is supported by existing soil maps and geostatistics.

2. Materials and Methods

The study area is located in Northern Italy, in the alluvial plains of Emilia Romagna (lat. 43°50'N-45°00'N; long. 9°20'E -12°40'E Greenwich approx). The area is about 12,000 km^2, with elevations between -3 and 150 m a.m.s.l. Climate is temperate-sub oceanic, with a mean air temperature of 12.4°C (max. and min. 19.3 and 8.2 °C respectively), and a mean annual precipitation ranging from 520 to 820 mm. Rainfall and potential evapotranspiration data, for the period 2004-2010, come form a network of about 300 pluviometric stations. The spatialisation of these data is obtained by ordinary kriging with a grid of 5x5 km. In the study area a net of 160 piezometer wells is present. Though the sites are monitored since 1997, the network has been set up in different monitoring projects carried on at provincial level. The temporal series are therefore highly heterogeneous, in terms of temporal continuity, frequency of observations, and spatial distribution. On average 1.7 readings per month are available at the sites. The predictive model was developed using Evolutionary Polynomial Regression (EPR, Giustolisi and Savic, 2006). EPR are regression-based algorithms, which use a hybrid between polynomial structures and evolutionary computing to model environmental processes and/or systems. The analysis was carried out using the EPR-Toolbox v.2.1.SA (Laucelli et al, 2010), using the normal SSE function as objective function. In order to calibrate the model, 47 sites were selected for a total of 5611 water table readings between 2004 and 2009. At each site, in correspondence with the readings, cumulate rainfall and potential evapotranspiration have been calculated at 30, 60, 180 and 365 days preceding the reading. From these values, evapotranspiration deficit has been calculated for the same time intervals. Furthermore, the average water table depth between 2004 and 2009 was calculated at each site (N = 118) and added to the list of predictive variables to be selected by the EPR model. The predictive model has been validated on a subset of independent data (N = 2052) from 15 sites not used for model calibration, with readings between 2005 and 2010.

3. Results

The descriptive statistics of the calibration data set (N = 5611) are reported in Table 1. The optimal expression provided by EPR to estimate the water table depth at any given day *i* has the following form:

$$WT_{depth\_i} = -0.0016163 * D_{180} * WT_{avg} + 0.00000064746 * P_{cum30} * P_{cum60}^{0.5} * WT_{avg} + -0.000074113 * ETP_{365}^{0.5} * P_{cum30}^{0.5} * P_{cum60} * WT_{avg}^{0.5} + 0.00099791 * ETP_{365} * WT_{avg} + +0.0000000045408 * ETP_{365} * WT_{avg}^{2} * D_{180} - 19.564$$

The EPR model was tested on an independent data set (N = 2052); calibration and validation results are shown in Table 2. Figure 1 shows the time series of observed vs. predicted values for one of the validation site.

	WT_{depth} (cm)	ETP_{365} (mm)	ETP_{180} (mm)	ETP_{90} (mm)	ETP_{60} (mm)	ETP_{30} (mm)	P_{cum365} (mm)	P_{cum180} (mm)	P_{cum90} (mm)	P_{cum60} (mm)	P_{cum30} (mm)	D_{180} (mm)
Mean	158	1031	508	292	203	106	573	276	139	95	49	-232
Min.	3	869	173	48	26	12	243	26	4	0	0	-762
Max.	300	1166	914	537	382	203	1173	857	723	494	312	631
Std. Dev.	63	47	223	149	104	55	138	87	68	57	40	245

Table 1. Descriptive statistics of the calibration data set (N = 7,050).

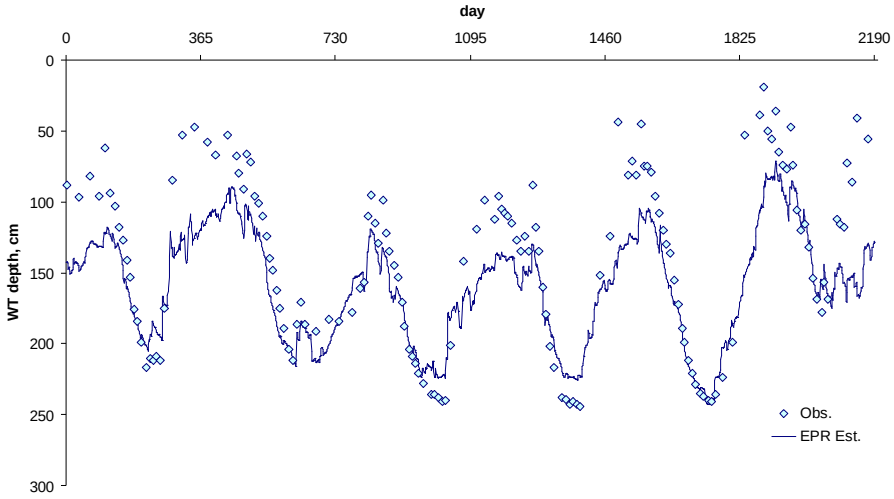


Figure 1: Estimated vs. observed water table depth at a validation site.

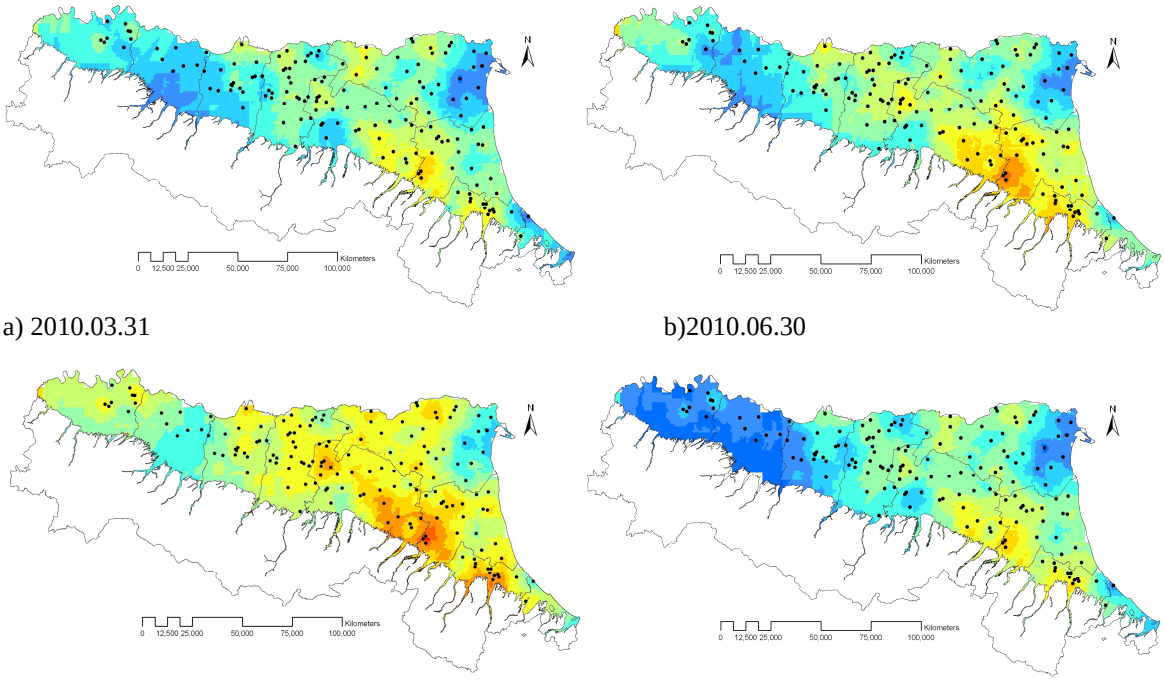
	Calibration data set (N= 5,611)	Validation data set (N=2,050)
R^2	0.71	0.65
RMSE, cm	34.14	37.68
MAE, cm	26.27	30.04
E', -	0.51	0.43
AIoA/d', -	0.73	0.69

Table 2. Model performance. MAE, mean absolute error; RMSE, root mean squared error; E', modified coefficient of model efficiency; IoA', modified index of agreement.

In terms of input data requirements, the EPR model relies on climatic data at any given site in the study area and on the long term average of WT depth. . In order to provide spatio-temporal map of soil water table depth at any given day *i*, the average WT depth has been estimated for the whole area via ordinary kriging over a 1x1 km grid using the available sites between 2004 and 2009 (N=118). The parameter of the model fitted to the experimental omnidirectional variogram are shown in table 3. Concerning the capability of the model in extrapolating in the spatial and temporal domains, Figure 2 shows the water table depths at four different dates in 2010, whose data were not used for model calibration. Water table levels show distinct geographic patterns across the plain, which are more evident in the dryer periods. For example, the recently reclaimed lowlands in the north-eastern part of the plain, where the water table level is hydraulically controlled, are characterized by shallower WT (110-150 cm) while the opposite is observed in the south-eastern portion of the study area, which is characterized by a deeper water table (>175 cm). Results are coherent with rainfall spatial distribution, whose relevance in affecting WT depth is variable in time and space.

c_0	c_1	c_2	a_1 (m)	a_2 (m)	IGF
0.38	0.35	0.31	14,400	44,000	1.68e-03

Table 3. Variogram model parameters: c_0, nugget variance; c_1, c_2, sill components, and a_1 and a_2 the correspondent ranges. IGF: Indicative Goodness of Fit (Pannatier, 1996).



a) 2010.03.31 b)2010.06.30

c) 2010.09.30 d) 2010.12.31

Figure 2: Estimated WT depth at four dates in 2010.

4. Concluding remarks

The seasonal occurrence of shallow groundwater as water source for agricultural crops is crucial for a sustainable use of water resources. The empirical regionalised model presented in this paper, based on daily climatic records and on spatial estimates of average water table depth provides a tool to predict and map water table depth. The model allows for projecting ahead WT depth for the whole area regardless the availability of new WT readings, using daily climatic records and taking explicitly into account the spatial variability of WT and the spatial and seasonal variability of rainfall.

References

Giustolisi O. & Savic D.A. (2006) A Symbolic Data-Driven Technique Based on Evolutionary Polynomial Regression, J. of Hydroinformatics, 8(3), 227-222 .

Laucelli, D., Berardi, L., Doglioni, A. 2010. EPR- Toolbox v. 2.1.SA. Technical University of Bari, Dpt. Of Civil and Environmental Engineering

Morgan, C.P., Stolt, M.H., 2004. A comparison of several approaches to monitor water-table fluctuation. Soil Sci. Soc. Am. J., 68, 562 – 566.

Pannatier, Y., 1996. Variowin: Software for Spatial Analysis in 2D, Springer, New York 91 pp.

Assessment and modelling of spatial variability of the soil factors potentially affecting groundwater nitrate contamination in two agricultural areas of Molise Region (Southern Italy)

Colombo[1] C., Palumbo[1] G.,
[1] Dipartimento di Scienze Animali, Vegetali e dell'Ambiente, Università del Molise,
Via De Sancits s/n 86100, Campobasso (CB) Italy colombo@unimol.it

Sollitto[2] D., Castrignanò[2] A.
[2] Consiglio per la Ricerca e la Sperimentazione in Agricoltura (CRA)
Unità di Ricerca per i Sistemi Colturali degli Ambienti Caldo-Aridi
Via Celso Ulpiani, 5 70125 Bari (BA) Italy

Abstract: In this study spatial variability was used to analyze soil factors influencing the occurrence of high nitrate concentrations in agricultural soils in the Molise region of Southern Italy. The proposed methodology applied to two agricultural areas combines measurements of soil nitrate concentrations carried out by a monitoring network of 164 top-soils. A multivariate approach based on multivariate geostatistics and GIS was used to model spatial variability of the soil variables. The maps of each individual soil variable and regionalised factor show the areas of the landscape that might cause nitrate loss from agriculture soils. The results can be used to support sustainable land use planning in order to mitigate soil nitrate leaching.

Keywords: nitrate contamination, soil variability, multivariate approach, geostatistics.

1. Introduction

Evaluation of nitrate loss from agricultural soils is a useful tool to support sustainable land use planning. An understanding of the spatial-temporal variability of important soil properties and associated nitrate contamination can provide a framework for assessing and modelling of the main processes occurring in the soil. Many factors may affect the spatial distribution of nitrate in the soil and the consequent nitrate pollution of groundwater. Important factors include topography, hydrogeology, climate, pedology, land use, and the type of crop (Power and Schepers, 1989). All of these factors need to be accounted for when analyzing spatial distribution of nitrate in soil. On the other hand, the spatial patterns of these factors do not change over short periods of time and, therefore, are not the major contributors to changes in spatial distribution of nitrate from year to year (Marriott et al., 1997). The aim of our research is to analyze nitrate concentrations in agricultural soils with respect to specific explanatory soil variables, using GIS and geostatistical methods to delineate areas at different risk of soil nitrate leaching as a result of soil management.

2. Materials and methods

The study sites were Campomarino and Venafro, two agriculturally fertile areas in the region of Molise (southern Italy). The land use consists mostly of olive orchards, vineyards, fruit orchards, maize, and horticultural crops. The town of Venafro is located near an important agricultural plain, identified as a nitrate vulnerable zone by the European Community according to the EU Nitrate Directive (91/976/EC). Campomarino is a small farming town in lower Molise. Molise has a typical Mediterranean climate, with mean annual rainfall varying from 600 mm to 1500 mm, and mean annual temperature ranging from 10 to 16° C. Surface soil samples (Ap horizons) were collected to a depth of 0.40 m. A total of 71 samples were collected in Venafro and 63 samples were collected in Campomarino. The variables analysed were: pH, texture, available water capacity (AWC), cation exchangeable capacity (CEC), electrical conductivity (EC), $CaCO_3$ content, total organic carbon (TOC), and total nitrogen (Ntot) according to the standard Soil Methods of Analysis. Nitrate-N was extracted from the field-moist soil samples with 0.1 M KCl solution at a soil:solution ratio of 1:2 and determined colorimetrically.
Statistical data analysis was done in two steps. First, classical descriptive statistics were determined, and then geostatistical analysis was performed to investigate spatial dependence, to map soil variables, and to delineate homogeneous areas.
Even if ordinary cokriging does not require the data to follow a normal distribution, variogram modelling is sensitive to strong departures from normality because a few exceptionally large values may contribute to many very large squared differences. To produce the map of the variables we used multi-Gaussian cokriging (Wackernagel, 2003). The multivariate spatial data were analysed by cokriging and Factor coKriging Analysis (FCKA). The theory underlying FCKA has been described in many papers (Castrignanò et al., 2000; Wackernagel, 2003). The approach consists of decomposing the set of original second-order random stationary variables into a set of reciprocally orthogonal regionalized factors, related to N_S spatial scales. The three basic steps of FKA are the following:
1) modelling the coregionalization of the set of variables using the so called Linear Model of Coregionalization (LMC) and interpolating the variables by cokriging;
2) analysing the correlation structure between the variables by applying Principal Component Analysis (PCA) at each spatial scale;
3) cokriging a set of specific factors at each characteristic spatial scale and mapping them.
All statistical and geostatistical analyses were done by using the software package ISATIS®, release 11.0.

3. Results

The spatial maps of the eight raw variables from the Venafro site, obtained by cokriging on a 10 m x 10 m square grid cell, display distinct spatial patterns and also reveal some degree of spatial association among the different textural attributes. The surveyed area can be roughly divided into two main zones of approximately equal extent along the NW-SE direction. The southern part is characterised by higher clay contents, while the northern part is coarser textured. The spatial maps of the ten raw

variables from Campomarino, obtained by cokriging on a 10m x 10m square cell-grid, were better structured spatially, probably due to the impact of topography and soil parent material. As regards the textural properties, the higher contents of sand occur along the sea coast up to a distance of 2500 m inland, whereas clay is more concentrated at the south-east and south-west corners. The maps of Ntot and TOC show a wide area characterized by higher values in the south-eastern part where the soils are manly finer textured. The map of N-NO_3 are more variable, probably due to the impact of agricultural management. There is a wide central inner area characterized by higher values, which means that this area is potentially at risk of contamination (in Figure 1 only some maps are shown).

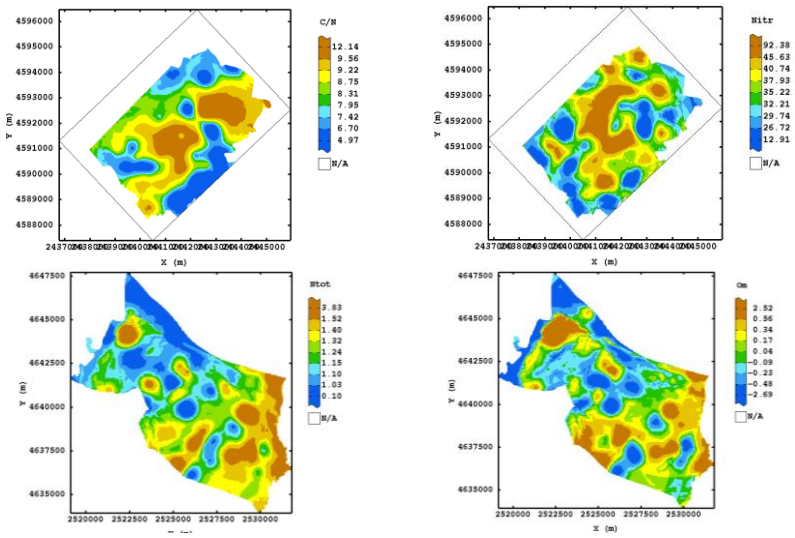
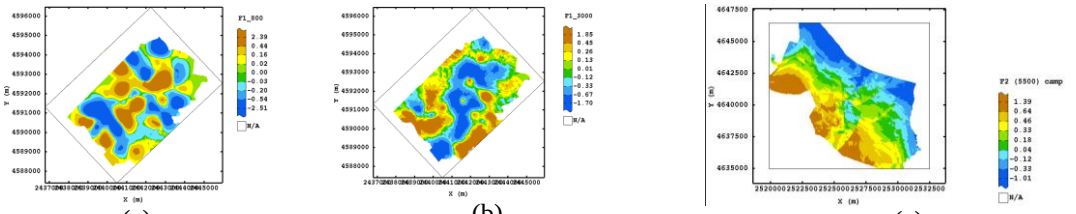


Figure 1. Spatial distribution of C/N ratio and N-NO_3 (Nitr) in Venafro soil (top) and Ntot and TOC (OM) in Campomarino soil.

To synthesize the complex, multivariate variation of the two areas in a small number of zones, to be ranked as to different risks of contamination, the factor cokriging analysis was applied separately to the two data sets from Venafro and Campomarino. The main component of variation for Venafro occurs within a range of 800 m, whereas for Campomarino the spatial variation is dominated by the structured components at both short (1000 m) and long (5500 m) ranges. In the following analysis we have retained only the eigenvectors producing eigenvalues greater than one and omitted the ones corresponding to nugget effect because the latter are mostly affected by measurement errors. Therefore, we focus for Venafro on the first factors at shorter (800 m) and longer scale (3000 m) which account for about 61% and 56%, respectively, of the variation at the corresponding spatial scales. For Campomarino the first two factors at shorter range (1000 m) and the first factor at longer range (5500 m) account for 45%, 25% and 78%, respectively, of the related spatial scale variation. The loading values for the factors (data not reported) indicate that for Venafro the TOC and clay content and, to a lesser extent, Ntot and C/N, as the most influencing first factor at shorter range. On the other hand, CSC and, to a lesser extent, silt content, TOC, C/N, and fine sand weigh more, but negatively, on the first factor at longer range. As for Campomarino, clay content and, to a lesser extent, N-NO_3 weigh more and positively on the first factor at shorter range, whereas TOC and Ntot weigh more on the second factor. The first factor at longer range is quite

exclusively dominated by elevation and partially and negatively by pH and $CaCO_3$. Figures 2 a-b show the maps of the two factors for Venafro. The one at short scale looks more variable, characterised by many spots of about 800 m wide with contrasting values of Ntot probably due to differences in land use and management. The soil factor at longer range is more related to CSC and partly to TOC and Ntot contents and looks better structured spatially.



(a) (b) (c)

Figure 2. Maps of the first regionalized factors at shorter and longer range (a-b) of Venafro and of the first regionalized factor at longer range (c) of Campomarino.

The map of the first factor at longer range (Fig. 2c) reproduces the topographic patterns faithfully and shows also a wide area characterized by higher values of $CaCO_3$ and pH at the north-west corner of the area.

4. Concluding remarks

In this study a multivariate geostatistical approach on different soil parameters was used to delineate the zones which might cause contamination by nitrate loss from agriculture soils. The resulting zones were also used to characterise spatial variability in physical and chemical soil properties that may potentially have an impact on soil nitrate contamination. In order to give useful site-specific recommendations to farmers for N fertilization, finer sampling is necessary.

References

Castrignanò A., Giugliarini L., Risaliti R., Martinelli N. (2000) Study of spatial relationships among some soil physico-chemical properties of a field in central Italy using multivariate Geostatistics. *Geoderma*, 97, 39-60.

Marriott C.A., Hudson G., Hamilton D., Neilson R., Boag B., Handley L.L., Wishart J., Scrimgeour C.M., Robinson D. (1997) Spatial variability of soil total C and N and their stable isotopes in an upland Scottish grassland. *Plant Soil*, 196, 151–162.

Power J. F., Schepers J. S. (1989) Nitrate Contamination of Groundwater in North America. Agriculture. *Ecosystems and Environment*, 26, 165–187.

Wackernagel H., (2003) *Multivariate Geostatistics: An Introduction with Application*. Springer Ed., New York, 387.

Webster R. (1985) Quantitative spatial analysis of soil in the field. *Advances in Soil Science*, vol. 3. Springer, New York, 1–70.

Assessment of spatial and temporal within-field soil variability by using geostatistical techniques[1]

Annamaria Castrignanò[1]
[1]CRA – SCA, Research Unit for Cropping Systems in Dry Environments
Giovanna Cucci[2]
[2] Department of Agri –Environmental and Land Sciences, University of Bari, Italy
Mariangela Diacono[1]
[1]CRA – SCA, Research Unit for Cropping Systems in Dry Environments, Via C. Ulpiani 5, 70125 Bari, Italy, mariangela.diacono@inwind.it
Daniela De Benedetto[1]
[1]CRA – SCA, Research Unit for Cropping Systems in Dry Environments
Giovanni Lacolla[2]
[2] Department of Agri –Environmental and Land Sciences, University of Bari, Italy
Antonio Troccoli[3]
[3]CRA – CER, Cereal Research Centre, Foggia, Italy

Abstract:
The main objective of this study was to analyze, with geostatistical techniques, the soil variability for some soil parameters on a 12-ha field cropped with durum wheat in Foggia (Southern Italy). Soil samples were collected at 100 georeferenced locations in 2005 and 2007. The application of multivariate geostatistical technique, called factorial co-kriging, allowed the delineation of the field into 3 main clusters. Contingency tables, k statistics and Q-Q plots were applied to assess the temporal variation.
The results showed a significant increase in soil organic matter and a decrease in P content up to 40 cm depth during the trial period.

Keywords: Soil variation; Multivariate geostatistics; Organic matter; Phosphorous.

1. Introduction

Soils commonly exhibit within-field spatial variability of some inherent properties such as texture, depth of topsoil and organic C content, resulting from complex geological and pedological processes acting over different spatial and temporal scales. Therefore, soil variables are expected to be correlated in a scale-dependent way (Castrignanò et al., 2000). Moreover, meteorological conditions and anthropogenic activities, such as tillage and fertilization, may cause spatial and temporal variation in soil (Basso et al., 2009).
The main objectives of this work were to characterize the soil variation of a field in southern Italy and test whether the soil management has affected chemical fertility significantly. We deemed a geostatistical analysis of coregionalization to be more revealing than a univariate approach, and we studied the scale-dependent correlation

[1]This research was funded by the Integrated system for development of southern cereal farming (SI.Cer.Me). Program for southern Italy development: Research and Technological Innovation. Resolution CIPE 17/2003 -1.1 and 83/2003.

structure of some soil properties, focusing on the delineation of the field into homogeneous areas.

2. Materials and Methods

The research was carried out in a 12-ha field cropped with durum wheat (*Triticum durum* Desf.) at the Experimental Farm of the CRA - Cereal Research Centre (Foggia, 41° 27' N, 15° 36' E, south-eastern Italy), during November 2005 – July 2008 period. The soil is silty-clay Vertisol of alluvial origin, classified as Fine, Mesic, Typic Chromoxerert by Soil Taxonomy-USDA. The climatic conditions were characteristic of a Mediterranean environment, with a drier season between May and September and cold likely returning in the spring months (March-April).
One hundred georeferenced locations were selected so to evenly cover the field. Soil samples were collected at these locations, before sowing and fertilization, to 0-40 cm in 2005 and to 0-20 and 20-40 cm in 2007, and analyzed for sand, silt and clay contents (%), organic matter (%, OM), available P (mg kg^{-1}, expressed as P_2O_5), according to the standard methods of soil analysis (Pagliai, 1997; Violante, 2000).
The multivariate spatial and temporal data set was jointly analyzed by cokriging to produce thematic maps and by the Factor Co-kriging Analysis, developed by Matheron (1982), to delineate homogeneous areas. The geostatistical analyses were performed with ISATIS (Geovariances, 2010). Contingency tables and k statistics were calculated to assess the spatial association of the P maps and the ones of OM at the two dates. Q-Q plots were used to test the temporal trend. The approach was implemented with the FREQ procedure of the SAS/STAT software package (SAS/STAT Software Release 9.2, 2010).

3. Results

The exploratory analysis of the data revealed considerable spatial variation in soil properties at each sampling date and most variables were significantly correlated, which justified the choice of a multivariate approach. A linear model of coregionalization (LMC) was fitted to all both direct and cross-variograms, including the following basic spatial structures: a nugget effect and two exponential models with a range of 100 and 300 m.
Figures 1(a-d) display the co-kriged maps of the available P and the OM contents at the two sampling dates. The P maps were characterized by great erraticity, with several hot spots evenly spread over the field. However, it is possible to notice a tendency to higher values on north-western border of the field at both dates. On the contrary, the OM maps seem better spatially structured and can be roughly split into three main zones: a southern part with generally higher contents, a central area with lower ones, and a more variable northern area. All maps showed a general consistency over time. However, to make less subjective the results of a visual inspection, the contingency tables for P and OM were calculated using a common classification in three isofrequency classes. The overall accuracy was 58 and 59% for P and OM, respectively. The results showed that the structures of spatial dependence for P and OM remained stationary over about 58% of the field, due to inherent soil variation, but also other dynamic factors affected their spatial distributions, more related to meteorological conditions and crop management. The Bowker's test of symmetry verified the significance of P and OM variation over

time and the simple kappa coefficient values were 0.38 and 0.39 for P and OM, respectively.

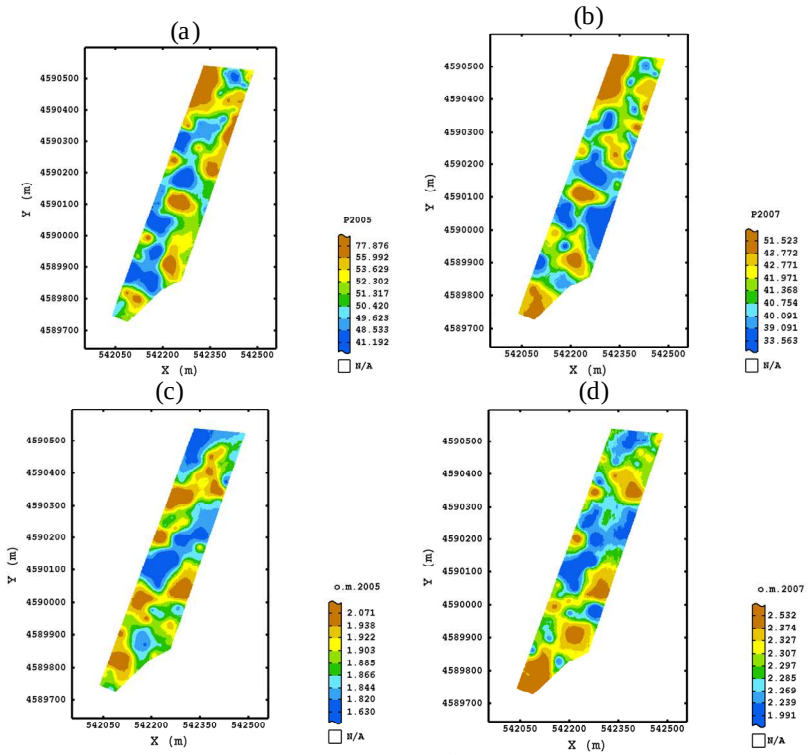


Figure 1: (a,b) Maps of available P (mg kg^{-1} P_2O_5) and (c,d) of organic matter (%) contents at the two dates

Actually, the spatial association between the two maps of either P or OM was not very high, confirming the dynamic character of soil fertility. However, the Q-Q plots (Figure 2) revealed a significant increase in organic matter and a significant decrease in P content during the trial period. Although a two-year trial period is not enough to draw general conclusions on observed P and OM trends, crop residue management very probably had a positive impact on increasing soil fertility.

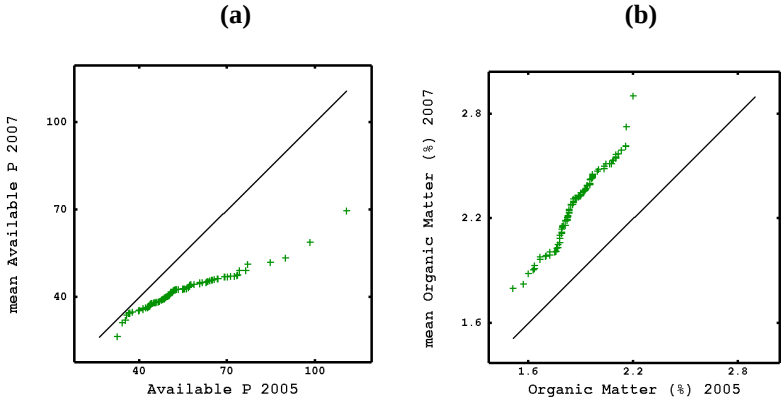


Figure 2: Q-Q plots of the variables P and OM for 2005 and 2007

As far as factor analysis, we retained only the first factor at longer range (=300 m; F1) with eigenvalue greater than 1, because it could produce a delineation of the field into

areas of size manageable by farmer. F1 was negatively correlated (data not shown) with OM and fine sand contents in both years and positively with clay contents and coarser sand content in both years, which also means that the main structures of spatial dependence are permanent over time.
The map of the first factor (Figure 3), displayed using three isofrequency classes, showed a wide central area with lower content of OM to 0-40 cm depth and higher contents of clay and coarser sand, whereas the northern and southern areas were characterized by higher content of finer sand and lower organic fertility. F1 could then be used as an indicator of soil organic fertility and soil texture.

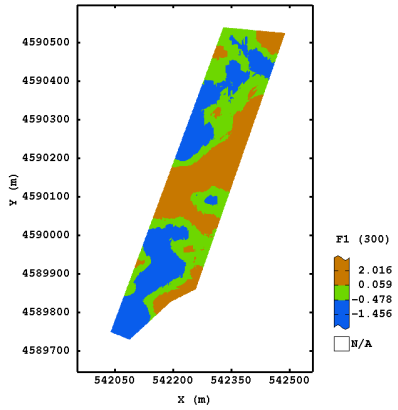


Figure 3: Map of the regionalized factor F1 produced using the Factorial co-Kriging procedure

4. Concluding remarks

Multivariate geostatistical analysis has produced the partition of an agricultural field into three homogeneous areas with different organic fertility and particle size distribution, which could be managed differentially. The results are encouraging and the approach might be used to test the effects of soil management over time.

References

Basso B., Cammarano D., Grace P.R., Cafiero G., Sartori L., Pisante M., Landi G., De Franchi S., Basso F. (2009) Criteria for selecting optimal nitrose fertilizer rates for precision agriculture, *Ital. J. Agron.* 4, 147-158.

Castrignanò A., Giugliarini L., Risaliti R., Martinelli N. (2000) Study of spatial relationships among some soil physico-chemical properties of a field in central Italy using multivariate geostatistics, *Geoderma*, 97, 39–60.

Geovariances, 2010. Isatis Technical Ref., release 10.04, Geovariances & Ecole Des Mines De Paris: Avon Cedex, France.

Matheron G. (1982) Pour une analyse krigeante des données régionalisées. Rapport N-732. Centre de Géostatistiques, École des Mines de Paris, Fontainebleau, France.

Pagliai M. (Ed.) (1997) Metodi di Analisi Fisica del Suolo (Physical Methods of Soil Analysis), Italian Ministry of Agriculture, Franco Angeli, Milan, Italy, (in Italian).

SAS Institute Inc. 2010. SAS/STAT Software Release 9.2, Cary, NC, USA.

Violante P. (Ed.) (2000) Metodi di Analisi Chimica del Suolo (Chemical Methods of Soil Analysis). Italian Ministry of Agriculture. Franco Angeli, Milan, Italy, (in Italian).

CYCAS-MED project: analysis at regional and local scale of climate change impacts on cereals yield in Morocco[1]

Antonella Bodini, Erika Entrade
Institute of Applied Mathematics and Information Technology
(CNR-IMATI, Milano, Italy), antonella.bodini@mi.imati.cnr.it

Carla Cesaraccio, Pierpaolo Duce, Pierpaolo Zara
Institute of Biometeorology, (CNR-IBIMET, Sassari, Italy)

Martin Dubrovský
Institute of Atmospheric Physics ASCR, Prague, Czech Republic

Abstract: The project CYCAS-MED: *Crop yield and climate change impacts on agriculture: adaptation strategies to desertification processes in the Mediterranean areas*, aimed at the development of tools and methodologies for the assessment of the response of three major crops in Morocco to climatic change. Results for durum wheat are here presented.

Keywords: climate change, cereals, crop growth models, land suitability, weather yield function

1. Introduction

The rising trend of global atmospheric carbon dioxide is expected to induce a change in climate. Despite the uncertainty regarding the magnitude of this climate change, assessments of its impacts on agricultural production are needed for both scientific and policy-making purposes. The complexity of climate-crop production interactions makes simulation a very useful and practical approach available for making the needed assessments (de Jong *et al*. 2003).
The CYCAS-MED project aimed at assessing the magnitude of the response of three major crops in Morocco (barley, soft wheat and durum wheat) to climate change. Weather daily data have been analyzed to outline the climate of Morocco. This analysis allowed the calibration of a stochastic weather generator that, in turn, has been used to provide future climate scenarios. At the regional scale, the relationship between weather and crop yield has been investigated by linear regression and, according to this relationship, the mean crop yield under different climate scenarios has been obtained. Moreover, the Land Suitability approach for rainfed wheat as been applied to analyse land use modifications. At the farm level, a further analysis has been carried out by applying the crop simulation model CERES-Wheat (Ritchie and Otter-Nacke, 1985).

[1] Project partially founded by the City of Milan under the program *Defending Biodiversity: Solidarity and International Cooperation*, Grants 2008

2. Materials and Methods

Data. Daily data on rainfall and temperature at 30 meteorological stations have been provided by the partner Insitut National de la Recherche Agronomique (Morocco) and refer to the period from 1973-2006. Annual crop yield data for 15 provinces come from The Ministry of Agriculture, Rural Development and Fisheries of Morocco.

Regional scale analysis: impact on annual production. The FAO methodology developed for Africa and based on the Crop Specific Soil Water Balance (CSSWB), and its implementation by the software AgroMetShell (AMS; Gommes 1993) have been used for operational crop yield forecasting. This model requires daily weather data only and is specific for regions lacking of more agronomical information. The output of AMS consists of several indexes: *water excess*, *water deficit* and *actual evapotranspiration* computed at several phenological phases, *water satisfaction index* (WSI) and *total water requirement*. The relationship between annual crop yield and these output variables (*weather yield function*) has been estimated by linear regression analysis. The stochastic weather generator M&Rfi (Dubrovský *et* al. 2004) has been calibrated on daily weather data and used to generate synthetic time series of current and future weather. Two different scenarios have been considered: SRES B1 (low impact) and SRES A2 (high impact; IPCC, 2000) and projections of climate at 2050 derived from the coupled atmosphere-ocean general circulation model HadCM3. The estimated regression equations have been used to compute the expected crop yield for each time series from current and future climate.

Regional scale analysis: impact on land use modification. Evaluating land suitability means defining the requirements of the different land-use types for each land units of a certain region. A land suitability classification for rainfed wheat growing in the province of Settat has been made. Land is classified as suitable, using 3 different classes of suitability, or not suitable.

Farm scale analysis. The crop simulation model CERES-Wheat (Ritchie and Otter-Nacke, 1985) included in DSSAT v. 4.0 (Jones *et al.*, 2003) has been used to predict grain yield of a local variety grown at six experimental farms for which all the pedological, climatic, genetic and agronomic information necessary for model calibration and evaluation are available.

3. Results

Climate analysis. An accurate analysis of the climate in Morocco is prevented by short and strongly discontinuous time series, especially for rainfall. Data from four stations only allow analyzing the inter-annual variability of temperatures. These data show a homogeneous increase of temperatures (Mann-Kendall test) and of the index Tn90, describing the number of warm nights (Frich *et al*. 2002). The other climatic indexes considered in this study do not show similar results (see Bodini *et al*. 2011).

Regional scale analysis: impact on annual production. The goodness-of-fit of the linear regression model was high ($R^2 > 0.85$), a part for a few cases. Regression analysis of annual crop yield on the AMS output variables highlights that WSI is always significant, as expected. However, a second significant variable is sometimes obtained, varying with place and cereal, whose meaning has to be further investigated. See Bodini *et al*. (2011) for detailed results.

The weather generator M&Rfi was able to well reproduce temperature data, however it strongly underestimated precipitation variability. However, as AMS allows a trade-off between amount of rainfall and length of the growing season, this underestimation do not seem to affect crop yield estimation. According to the weather yield functions estimated from the available data, mean future yield for each time series from the two scenarios have been computed and compared to those obtained from simulations representing the current climate. For durum wheat (Figure 2), crop yield will decrease everywhere, and in some places, like Marrakech (MAR) crop yield could halve. Similar results are obtained for the other cereals (see Bodini *et al*. 2011).

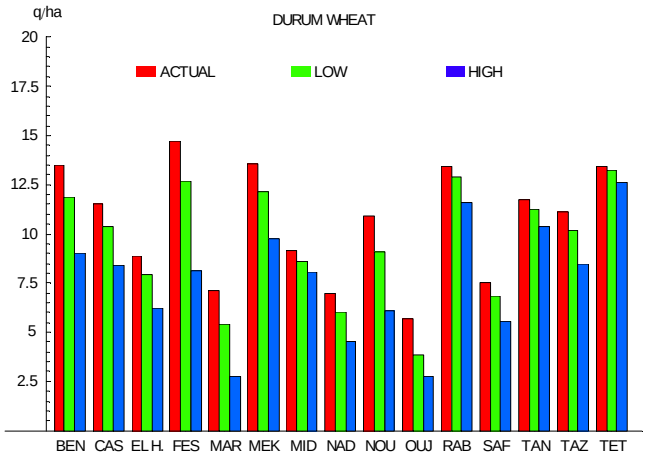


Figure 2 Mean expected annual crop yield for durum wheat at 2050, according to current climate (actual), scenario SRES B1 (low impact) and scenario SRES A2 (high impact). The plotted values are mean from 1000 simulations.

Regional scale analysis: impact on land use modification. Land Suitability analysis shows that the class of highest suitability (S1) for rainfed wheat became half at 2050 (high impact scenarios), and completely disappears at the 2100 projection, whereas S2 class reduces from 10% to 40%. In synthesis, a general reduction of major portions of territory suitability is evident, and the marginal suitability class (S3) increases in importance, doubling (2050 low and high, 2100 low) or tripling its incidence (high scenario 2100), as shown in Figure 3.

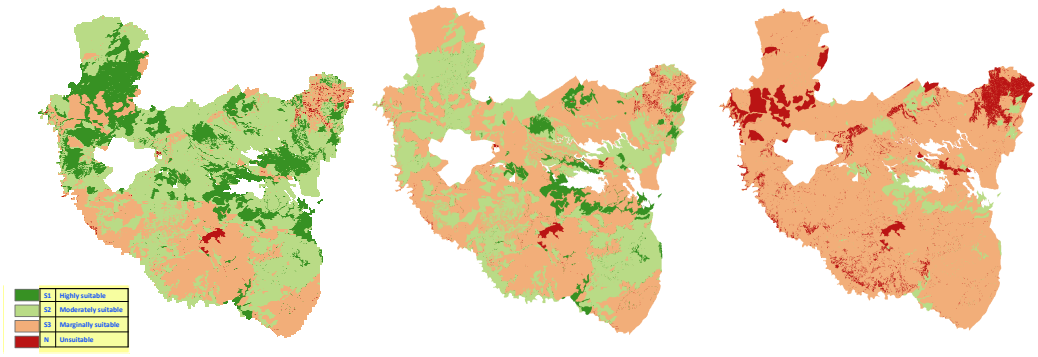


Figure 3 Map of Land Suitability for rainfed wheat in Settat province for actual climate (left), high impact scenario at 2050 (centre), and high impact scenario at 2100 (right).

Farm scale analysis. The application of the crop model at different locations shows a general tendency to a reduction of wheat yield moving from actual conditions to higher impact future scenarios. See Cesaraccio *et al*. (2011) for more details about this analysis.

4. Concluding remarks

The analysis of the impacts of future climate change scenarios highlights a significant reduction of the suitable areas for agriculture in Morocco and a significant reduction of rainfed cereals yield regardless of emission scenarios.
Adaptation strategies for responding to changes in climate regimes need to be investigated to adapt agricultural systems to the new conditions. From this perspective, tools and methods used in this project can be used to investigate other crops performances under changed conditions. Future work will concern the improvement of the WG and the investigation of other crop performances under climate change scenarios.

References

Bodini A., Cesaraccio C., Duce P., Entrade E., Zara P. (2011) Impatto dei cambiamen-ti climatici sulle produzioni agricole: strategie di adattamento ai processi di desertifi-cazione nelle aree mediterranee (CYCAS-MED). Report finale - Parte A, Technical Report IMATI-MI 11-1, Milano. Available on-line at the web site http://www.mi.imati.cnr.it/iami/abstracts/11-01.html

Cesaraccio C., Duce P., Zara P., Ferrara R., Pintus G., Bodini A., Dubrovský M. (2011) Impatto dei cambiamenti climatici sulle produzioni agricole: strategie di adattamento ai processi di desertificazione nelle aree mediterranee (CYCAS-MED), Report finale - Parte B. Technical Report IMATI-MI 11-2, Milano. Available on-line at the web site http://www.mi.imati.cnr.it/iami/abstracts/11-02.html

de Jong R, Li K.Y., Bootsma A., Huffman T., Roloff G., Gameda S. (2003) *Crop Yield and Variability under Climate Change and Adaptative Crop Management Scenarios*, Report, Climate Change Action Fund.

Dubrovský M., Buchtele .J, Zalud Z. (2004) High-Frequency and Low-Frequency Variability in Stochastic Daily W eather Generator and Its Effect on Agricultural and Hydrologic Modelling, *Climatic Change*, 63, 145-179.

Frich P., Alexander L.V., Della-Marta P., Gleason B., Haylock M., Klein Tank A., Peterson T. (2002) Global changes in climatic extremes during the 2nd half of the 20th century, *Climate Research*, 19, 193-212.

Gommes R. (1993) FAOINDEX, Version 2.1. Agrometeorology Group. FAO Rome.

IPCC (2000) *Emissions Scenarios. A Special Report of Working Group II of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge.

Jones J.W., Hoogenboom G., Porter C., Boote K., Batchelor W., Hunt L.A., Singh U., Gijsman A., Ritchie, J. (2003) The DSSAT cropping system model, *European Journal of Agronomy*, 18, 235-265.

Ritchie J.T., Otter-Nacke S. (1985) Description and performance of CERES-Wheat: use-oriented wheat yield model, in: *ARS wheat yield project*, ARS-38 National Technical Information Service, Springfield, VA, 159-175.

Geostatistical analysis and mapping of hydrocarbon pollutants in soils.

Chantal de Fouquet

Centre de Géosciences / Equipe Géostatistique, Ecole des Mines de Paris, France; e-mail: chantal.de_fouquet@ensmp.fr

Abstract: Data collected during sampling of the soils of former industrial waste lands are rarely scrutinized closely. However, exploratory analysis is an essential stage in order to describe the main characteristics of the concentration: possible heterogeneities, vertical variations, etc. Then the variographic analysis aim to characterize and to quantify the spatial variability. For hydrocarbon pollution, the very large spatial variability at small distances results in large uncertainties in the estimates. The kriged concentration map can be combined with the associated kriging standard deviation map to identify areas in which the uncertainties make it impossible to decide whether concentrations are greater or lower than a fixed quality threshold.

Examples are given using data from sites polluted by hydrocarbons.

Geostatistical analysis of groundwater nitrates distribution in the Plain d'Alsace

Rose-Line Spacagna
Università degli Studi di Cassino, Facoltà di Ingegneria - DIMSAT,
rlspacagna@unicas.it

Chantal de Fouquet
Mines ParisTech (Ecole des Mines de Paris), Géosciences - Géostatistique

Giacomo Russo
Università degli Studi di Cassino, Facoltà di Ingegneria - DIMSAT

Abstract: The groundwater of the Plaine d'Alsace (France) is one of the largest water reservoir in Europe. It is highly vulnerable to pollutants coming from anthropic activities (industrial and agricultural). The monitoring carried on by the local agencies on the groundwater pollution concurred to form a very large public database on water quality. In this study the nitrate concentration has been examined in detail for two years, namely 1997 and 2003, for which dense sampling was available, and the evolution of the distribution of the pollutant has been highlighted.

Keywords: groundwater, nitrates, geostatistics, estimations

1. Introduction

The cartographic representation of a pollutant distribution in a large area is a key tool in the safeguard of the hydraulic resources. From the monitoring data, a correct description of the pollution can be reached taking account of the regional character of the relevant variables (G. Matheron, 1972), in the geostatistical framework (de Fouquet, 2006). The groundwater of the Plaine d'Alsace in the North East of France is highly vulnerable to pollutants, due to the intense anthropic activities (agricultural, industrial, etc.) on the whole area (BRGM, 2006). The use of the hydraulic resource for domestic and industrial purposes makes its safeguard essential for the sustainable development of the region. The present study examines the distribution of nitrates and its evolution for the years 1997 and 2003.
Quantitative and qualitative information about the groundwater were deduced form ADES groundwater national portal. D'Agostino et al. (1998) analyzed the distribution of nitrates in the groundwater of Lucca plain (Italy) with reference to three different periods of the same year. In this study the concentration of nitrates measured in the 1997 and 2003 were analysed with reference to the same time interval, namely August and September of each year. A multivariate approach allowed the representation of the evolution of nitrates concentration in the reference time interval.

2. Data Analysis

Figure 1 shows the nitrates concentration monitoring points for 1997 and 2003 data sets (i.e. 696 and 601 monitoring points respectively for 1997 and 2003). Full details on data sets can be found in Spacagna (2009). In order to perform a multivariate variographic analysis, 574 monitoring points belonging to both data sets were identified. The considered measures are related to a seven weeks time interval, during which stationary hydraulic regime for groundwater was assessed. In Table 1 the main statistical parameters of the two data sets are summarized. The histograms highlight the frequencies of concentrations over the threshold value (red for 1997 and blue for 2003 in Figure 1). The highest concentrations of nitrates are mostly located along the western border of the geographical area, between the Vosges and the Plain of Alsace, characterized by geological discontinuities, relatively small thickness of the groundwater and intensive agricultural activities.

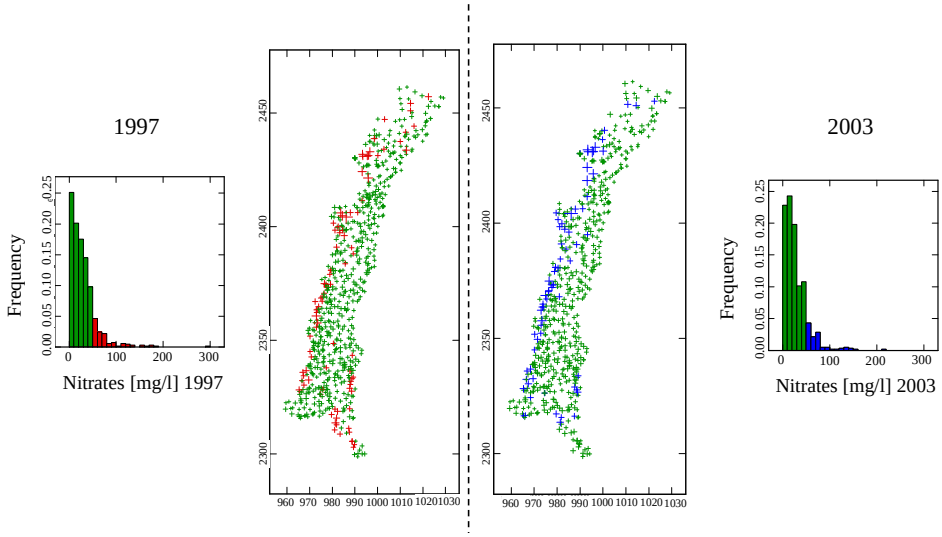


Figure 1: Monitoring points localisation a)1997; b) 2003 and nitrates concentration histograms

The good correlation ($\rho = 0.86$) between data of 1997 and 2003 sets is shown by Figure 2, where the nitrate concentrations are grouped in 30 classes ranging from 0 to 300 mg/l. The values greater than 100 mg/l or between 50 and 100 mg/l are mainly located under the bisector line, as well as the empirical regression of 2003-concentrations on 1997-concentrations. This suggests a decrease in the average concentration from 1997 to 2003.

Directional variographic analysis on different observation scale highlighted a geometrical anisotropy of the concentration at great distances, with a greater continuity along the N15° direction, according to the main flow direction of the Rhin river, whereas no anisotropy was detected at small distances (up to 5 km). The experimental variogram was then calculated with reference to the directions 15°, 60°, 105° and 150° setting a 4 km step. The simple variograms for 1997 data (γ_{97}) and 2003 data (γ_{03}) and the cross-variogram ($\gamma_{97,03}$) are fitted by means of a co-regionalisation linear model, considering the phenomenon as a sum or superposition of independent processes occurring at different spatial scales. The parameters of the theoretical variogram models are reported in Table 2.

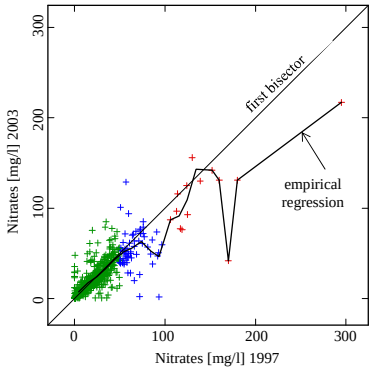


Figure 2: Correlation between the concentration of nitrates in 1997 and 2003

	1997		2003	
	complete set	common data	complete set	common data
points	696	574	601	574
minimum [mg/l]	0.17	0.17	0,50	0.50
maximum [mg/l]	295	217	217	217
average [mg/l]	28.06	27.28	27.36	27.30
std. dev. [mg/l]	23.00	24.60	24.67	24.39
variation coeff.	0.98	0.90	0.90	0.89
kurtosis	17.83	19.34	9.93	9.94
skewness	3.11	3.20	2.39	2.35

Table 1: 1997 and 2003 nitrates concentration data set.

	model	**sill**	**Range**
γ_{97}	nugget effect	166	-
	isotropic spherical	360	5 km
	anisotropic spherical	280	N15: 65 km / N105 15 km
γ_{03}	nugget effect	255	-
	isotropic spherical	140	5 km
	anisotropic spherical	248	N15: 65 km / N105: 15 km
$\gamma_{97,03}$	nugget effect	162	-
	isotropic spherical	220	5 km
	anisotropic spherical	242	N15: 65 km / N105: 15 km

Table 2: Simple and cross variograms model parameters

3. Estimation

In order to estimate the evolution of nitrates concentrations during the selected time interval, the difference D between 1997 and 2003 concentrations at the same measure points is introduced. The simple and cross variograms (namely γ_D and $\gamma_{97,D}$) of the difference D and the 1997 concentration are derived from the concentration bivariate model :

$$\gamma_D(h) = \gamma_{97}(h) + \gamma_{03}(h) - 2\,\gamma_{97,03}(h) \text{ and } \gamma_{97,D}(h) = \gamma_{97,03}(h) - \gamma_{97}(h);$$

the models parameters are summarised in Table 3. As the cokriging ensure the consistency between the estimations of different variables, it is equivalent to cokrige the two concentrations and to calculate their difference or directly to cokrige the difference from the two concentrations data (Rivoirard, 2003).

Figure 3a shows the nitrate concentration in 1997 estimated by means of cokriging. Figure 3b represents the cokriging of D with 1997, with evidenced the isofrequency classes of variation, whereas in Figure 3c the standard deviation of the estimation error of cokriging is presented.

	model	sill	Range
γ_D	nugget effect	97	-
	isotropic spherical	60	5 km
	anisotropic spherical	44	N15: 65 km / N105: 15 km
$\gamma_{97,D}$	nugget effect	-4	-
	isotropic spherical	-140	5 km
	anisotropic spherical	-38	N15: 65 km / N105: 15 km

Table 3: Simple and cross variograms model parameters

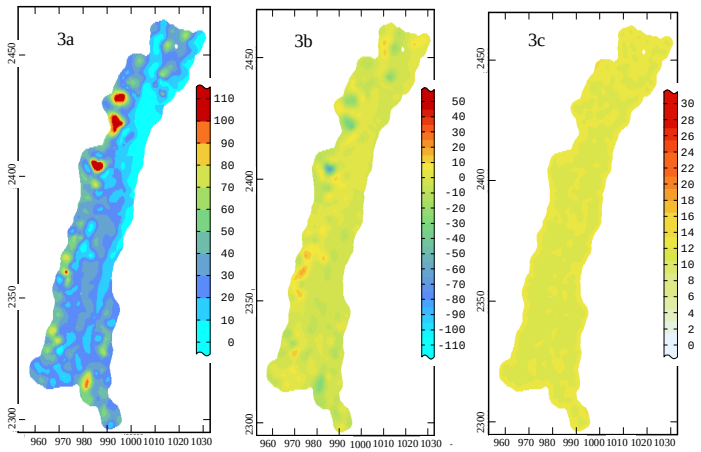


Figure 3. Estimation of nitrates concentration and its evolution between 1997 and 2003: a) 1997cokriging; b) Cokriging of the difference 2003-1997; c) standard deviation of estimation errors of the cokriged difference

4. Concluding remarks

The spatial structure of the nitrates concentrations was investigated by means of multivariate variography, highlighting the anisotropy closely related to the prevailing direction of groundwater flow (APRONA, 1999). Based on the correlation between the data collected at the same monitoring points and on the study of the difference of nitrates concentrations, a reduction of the highest concentrations of nitrate between 1997 and 2003 data was observed.

References

APRONA, 1999. Carte des directions d'écoulement de la nappe de la Plaine d'Alsace, septembre 1999, www.aprona.net.

BRGM, 2006. Aquifères et eaux souterraine de France. Collection Scientifique et Technique.

D'Agostino V., Greene E.A., Passarella G., Vurro M., 1998. Spatial and temporal study of nitrate concentration in groundwater by means of coregionalization. Environmental Geology 36 (3-4) December 1998. Springer-Verlag.

de Fouquet C., 2006. La modélisation géostatistique des milieux anthropisés. Memoire des Sciences de la Terre N°2006-13. Academie de Paris, Université Pierre et Marie Curie.

Matheron G., 1972. The theory of regionalized variable and its applications. Les Cahiers du Centre de Morphologie Mathématique. Ecole des Mines de Paris, Fontainebleau.

Rivoirard J., 2003. Course on multivariate geostatitics. C-173. Centre de Géostatistique de Fontainebleau. Ecole des Mines de Paris.

Spacagna R-L, 2009. Evoluzione della concentrazione di nitrati nella falda della Piana d'Alsazia. Master Degree in Civil Engineering. University of Cassino, Italy.

Influence of different olive grove management on spider diversity

Pamela Loverre, Rocco Addante
Dipartimento di Biologia e Chimica Agro-Forestale ed Ambientale, Università degli Studi di Bari "Aldo Moro", Italy,
pamela.loverre@hotmail.it

Crescenza Calculli
Dipartimento di Scienze Statistiche "Carlo Cecchi", Università degli Studi di Bari "Aldo Moro", Italy

Abstract: Spiders are known as one of the most important group of predators in olive agroecosystems, limiting the populations of insect pests and the damage they can cause. However, some agricultural practices are known to modify spider communities changing their species composition and abundance. To assess the influence of different management systems we collected data on spider fauna in three different olive groves and with three different methods. For the three sampling methods biodiversity indexes as Simpson's, Shannon's and Sørensen's were calculated in terms of spiders' families, in order to evaluate their temporal evolution and the relation to crop management systems. This purpose is accomplished in the context of generalized linear models and cluster analysis of dissimilarity matrices.

Keywords: Araneae, biodiversity indexes, cluster analysis, cultural practices, dissimilarity coefficients, GLM, *Olea europaea*

1 Introduction

All spiders (Araneae) are predators that feed primarily on insects and other arthropods (Wise, 1993). Many studies have revealed that spiders are a large fraction of the predator fauna in agroecosystems, both in terms of population density and in diversity of species (Ghavami, 2006), representing the most diversified group and, after the ants, the most abundant group of predators in olive groves (Morris *et al.*, 1999). It was observed that spiders are more sensitive than their prey to pesticides: thus the absence of these predators can induce pest outbreaks (Maloney *et al.*, 2003). Some cultural practices, as the use of pesticides, bring changes in spider composition (Santos *et al.*, 2007). The purpose of the present study was to characterize spider biodiversity of three olive groves subjected to different intensities of cultivation practices. The effect of three different sampling tools was also evaluated.

2 Materials and Methods

In May 2010 a research was initiated in three olive groves in the countryside of Valenzano (Bari, Italy) to assess the influence of different management systems on the spider fauna. The study was conducted until March 2011 in three olive groves: a private olive grove (field A) and two experimental groves (fields B and C) managed by the Faculty of Agriculture (University of Bari). Field A was abandoned, at least over the last decade, Field B was under minimized agronomical practices and Field C was under a larger number of farming practices (insecticide treatment and weed control). Spiders were collected fortnightly using three different sampling methods: 1) pitfall traps for collecting wandering spiders at ground level; 2) cardboard bands placed around the trunk for spiders sheltering between the bark anfractuosities; 3) frappage for sampling spiders living on the foliage of the olive trees. In each grove, five pitfall traps about 15 meters apart were placed. Pitfall traps, having a diameter of 12 cm and a height of 12 cm, were buried up to the top and filled to fourth with a mixture to preserve the animals collected. A cardboard band about 15 cm high, was wrapped in 3 to 4 laps around the trunk of five olive trees per field, at about one meter above the ground level. Collection by frappage was carried out on five trees per grove, selected randomly at each sampling. Two branches per tree were beaten over an entomological umbrella (1m x 1m), collecting all the spiders dislodged. Overall, five units were taken per each sampling method (pitfall, cardboard bands, frappage on plant) and olive grove (A, B, C). In the laboratory all the spiders collected were identified using dichotomous keys. Most of the spiders were released in the respective collection field after identification. As measures of α-biodiversity the Shannon and Simpson indexes were calculated for each date, field and sampling method (summarizing the five replications). While the Shannon index depends on the number of families identified and on the evenness of their abundance, the Simpson index measures the probability that two individuals randomly selected from a sample will belong to the same family. The variation of both measurements with time, habitat and sampling method was investigated by generalized linear models (GLM, Zuur *et al.*, 2009). In order to compare the different habitats and sampling methods a measure of β-biodiversity as Søresen index was also considered. This index measures the dissimilarity between pairs of objects and was calculated for each combination of habitats and sampling methods. Cluster analysis based on such a dissimilarity matrix was subsequently applied.

3 Results

GLM's for the Shannon and Simpson indexes were fitted considering the same set of effects: habitat, sampling method and linear time trend. Standard routines contained in the statistical environment `R` (R Development Core Team, 2008) were used throughout. The two response variable were both preliminary transformed in order to obtain tractable marginal distributions. The monotone transformation $1/(1+x)$

Effects	Shannon			Simpson		
	estimate	SE	p-value	estimate	SE	p-value
Field A	2.390	0.103	< 0.000	1.573	0.049	< 0.000
Field B	2.573	0.107	< 0.000	1.613	0.049	< 0.000
Field C	2.769	0.110	< 0.000	1.740	0.051	< 0.000
bands	-	-	-	0.037	0.044	0.402
beat	-	-	-	0.111	0.043	0.012
time trend	-0.036	0.007	< 0.000	-0.011	0.003	< 0.000

Table 1: GLM's for Shannon and Simpson indexes, parameter estimates.

allows for null values of the biodiversity indexes (two observations with only one spider family) and produces a switch from left to right skewness compatible with the Gamma distributional assumption (inverse link). Overall model and effects significance were used to select relevant covariates of the two models reported in Tab. 1. The highest number of spiders was collected in field B, the lowest (less than half) in field C (the one with a higher cultural pressure), while intermediate values were observed in A. Gnaphosidae, mainly collected by cardboard bands, were the family clearly dominant in the three olive groves, they were followed by Zodariidae, collected exclusively by pitfall traps. The highest values of both Shannon and Simpson indexes were observed in the olive grove subjected to more intensive cultivation practices (field C), as a consequence of the greater evenness of spider families. While both indexes are significantly influenced by the habitat, only the Simpson index shows a significant difference of the frappage (beat) sampling method with respect to the other two. A negative linear time trend was detected for the two transformed indexes, implying a slight decrease of the α-biodiversity across time.

In Fig. 1 the heat map relative to the Søresen index (Borcard *et al.*, 2011) highlights similarities between habitats for each sampling method. For the frappage sampling method, the intensive management in field C can be distinguished from the others. From a cluster analysis point of view this leads to a higher heterogeneity in relation with this sampling method.

4 Concluding remarks

The higher intensity of some cultural practices, in particular treatment with insecticides, caused a decrement in biodiversity of spiders on the foliage of olive trees soon after the treatments. To obtain a more complete description of the spider fauna of an agroecosystem as the olive grove, different collection methods should be used simultaneously. The three methods used in the course of this research seem to complement each other, allowing to detect a greater number of families.

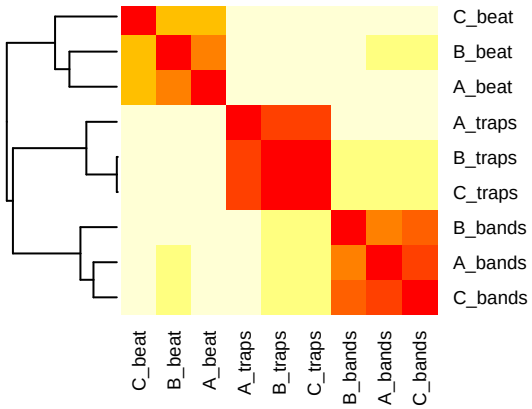


Figure 1: Heat map of the distance matrix based on Sørensen's index reordered according to the dendrogram.

References

Borcard D., Gillet F., Legendre P. (2011) *Numerical Ecology with* `R`, Springer, New York.

Ghavami S. (2006) Abundance of spiders (Arachnida: Araneae) in olive orchards in Northern part of Iran, *Pakistan Journal of Biological Science*, 9 (5), 795-799.

Maloney D., Drummond F.A., Alford R. (2003) Spider predation in agroecosystems: can spiders effectively control pest populations? *MAFES Technical Bulletin*, 190, 1-32.

Morris T.I., Symondson W.O.C., Kidd N.A.C., Campos M. (1999) Las arañas y su incidencia sobre *Prays oleae* en el olivar, *Boletin de Sanidad Vegetal Plagas*, 25, 475-489.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`.

Santos S.A.P., Pereira J.A., Torres L.M., Nogueira A.J.A. (2007) Evaluation of the effect, on canopy arthropods, of two agricultural management systems to control pests in olive groves from north-east of Portugal, *Chemosphere*, 67, 131-139.

Wise D.H. (1993) *Spider in ecological webs*, Cambridge University Press, Cambridge.

Zuur A.F., Ieno E.N., Walker N.J., Saveliev A.A., Smith G.M. (2009) *Mixed effects models and extensions in Ecology with* `R`, Springer, New York.

Landcover classification of agricultural sites using multi-temporal COSMO-SkyMed data[1]

G. Satalino, A. Balenzano, A. Belmonte, F. Mattia
CNR-ISSIA, Bari, Italy, satalino@ba.issia.cnr.it

D. Impedovo
Dipartimento di Informatica, Università degli Studi di Bari

Abstract: The objective of this paper is to report on the crop classification activities carried out during the first year of the Italian project "Use of COSMO-SkyMed data for LANDcover classification and surface parameters retrieval over agricultural sites" (COSMOLAND), funded by the Italian Space Agency. The project intends to contribute to the COSMO-SkyMed mission objectives in the agriculture and hydrology application domains. In particular, the objective of the classification activities is to assess the potential of multi-temporal series of X-band COSMO-SkyMed SAR data for crop classification. The selected agricultural site is located in the Capitanata plain close to the Foggia town (Puglia region, Southern Italy). Over this area, 8 Stripmap PingPong COSMO Sky-Med images at HH/HV polarization and at low incidence angle were acquired from April to August 2010. In the paper, a classification scheme based on the Maximum Likelihood algorithm is applied to the multi-temporal data set and its accuracy is assessed with respect to a reference map obtained by means of SPOT data.

Keywords: Land cover classification, SAR, COSMO-SkyMed, multi-temporal data

1. Introduction

The mapping of land cover/use and the monitoring of spatial and temporal variability of land surface parameters are important issues in the management of land and water resources. The improved spatial resolution and the reduced revisiting time of the new generation of spaceborne SAR systems, such as Cosmo-SkyMed, aroused an increasing interest in SAR data for land use classification. Several past studies have assessed the sensitivity of SAR data at C and L band to various crop or land classes and their use for crop mapping or land classification (McNairn et al., 2004, Skriver et al., 2010). On the contrary, relatively little work has been conducted up to date by using X-band SAR data due to the lack of long series of data. Nowadays, the availability of spaceborne SAR systems operating at X-band and characterized by a short revisiting time represents a good opportunity to deeper explore the use of this frequency for land cover classification.
In this context, the objective of this paper is to report on the crop classification activities carried out during the first year of the Italian project "Use of COSMO-SkyMed data for LANDcover classification and surface parameters retrieval over agricultural sites"

[1] This research is supported by the Italian Space Agency under contract I/051/09/0. COSMO-SkyMed data were provided by ©ASI in the framework of ©CSK AO 2161. SPOT data were obtained from CNES (2010) Distribution Spot Image ISIS-368.

(COSMOLAND), funded by the Italian Space Agency. The activities are still in progress and their final aim is to assess the potential of multi-temporal series of X-band COSMO-SkyMed SAR data (ASI ref., 2010) for crop classification.
In this paper, a multi-temporal series of X-band COSMO-SkyMed SAR data acquired in 2010 over an agricultural area in the Capitanata plain, Southern Italy, is investigated. The experimental data set is described in the next section, then the adopted classification algorithm and the first obtained results are illustrated and discussed. Finally, a summary and future work are drawn.

2. Materials and Methods

The investigated site (Figure 1) is located in the Capitanata plain close to Foggia town (Puglia region, Southern Italy), which is the second largest plain in Italy. The study area of approximately 700km^2 is mainly devoted to durum wheat cultivation (more than 50% of the total cultivated area). Other important seasonal crops are tomato and sugar beet. The classification image reported in Figure 1 shows an updated land cover map derived from 2 SPOT images acquired in 2010 (SPOT4 on 04/07/2010, and SPOT5 on 26/07/2010) classified by using the Maximum Likelihood algorithm. It is worth noting that wheat fields are already harvested in June, and therefore the class "wheat" on these dates was obtained classifying "harvested wheat". The overall accuracy obtained is approximately 89%.

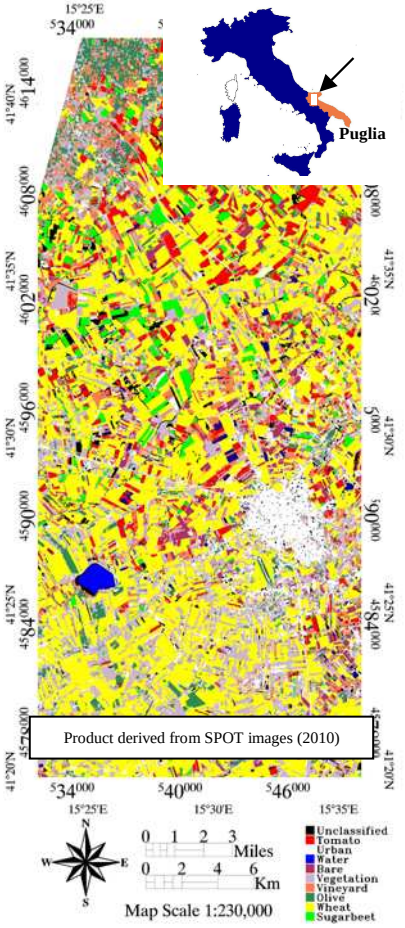


Figure 1: Land cover map of the Foggia site derived from SPOT data acquired in 2010.

Over this area, 8 Strip-map (Ping-Pong) level 1C-Geocoded Ellipsoid Corrected (GEC) products, at HH and HV polarization and at 26° mean incidence angle were also acquired from April to August 2010 (Table 1). It may be worth noting that there were no SAR acquisitions in June, which is an important period of the growing season.
The SAR images were coregistered, geocoded, and spatially filtered with a Boxcar filter of 5x5 pixels.

ID	Date	Mode swath	Mean incidence angle [°]	Polarization
D1	03/04/10	StripMap PP02	26	HH/HV
D2	27/04/10	StripMap PP02	26	HH/HV
D3	21/05/10	StripMap PP02	26	HH/HV
D4	29/05/10	StripMap PP02	26	HH/HV
D5	08/07/10	StripMap PP02	26	HH/HV
D6	24/07/10	StripMap PP02	26	HH/HV
D7	01/08/10	StripMap PP02	26	HH/HV
D8	09/08/10	StripMap PP02	26	HH/HV

Table 1: COSMO-SkyMed images (level 1C-Geocoded Ellipsoid Corrected (GEC) products) acquired over the Foggia test site in 2010 .

3. Results

Previous results have shown that single-date SAR data usually are not sufficient to accurately discriminate crop classes, on the contrary multi-temporal information can significantly improve the classification accuracy (Skriver et al., 2011). To investigate the extension to which such a result holds for the Foggia site, two different scenarios were compared:

1) classification of 1-date SAR image at HH & HV;
2) classification of 3-dates SAR images at HH & HV.

The selected classes are: wheat, sugar beet, tomato, vineyard and olives. Whereas, the selected dates are from ID D1 to D6, as reported in Table 1 (i.e. 6 out of 8 COSMO-SkyMed data-taken available), because they cover the main phenological cycles from April to July of the non-permanent crops. In August and September all the crops are either already harvested or about to be harvested. The adopted classification algorithm is the Maximum Likelihood (ML) for multivariate Gaussian distributed data, as SAR data with a number of looks larger than 10 can be assumed. Training data extracted from the SAR images in correspondence of fields cultivated with the selected crops, were identified and used in the ML algorithm.
The overall accuracy (OA%) of correct classification is reported in Figure 2 for the two investigated scenarios. By using single-date SAR data (i.e. D1,…,D6), the OA ranges between 70% and 80% whereas, by using multi-date SAR data (e.g. 3D4 means 3 dates: D4, D5 and D6), the OA ranges between 80% and 90%. Therefore, multi-temporal information can bring an improvement ranging between 10% and 20%. It is also worth emphasising that a significant dependence of the OA on the specific dates is observed. For instance, on D3 and D4 the OA is significantly lower than on D1, D2, D5 and D6.

This is likely due to the fact that on D3 and D4 there is a reduced separation in the radar response of the five crop classes related to their phenological stages.
Figure 3 shows an example of land cover image obtained from multi-temporal, dual polarization COSMO SkyMed images (i.e. case 3D4, HH+HV).

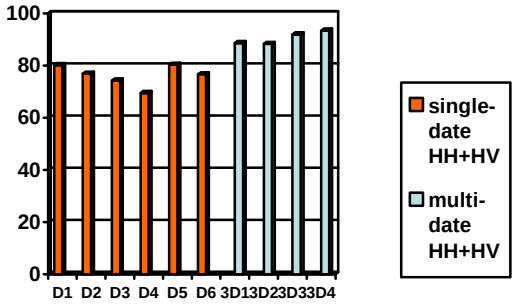


Figure 2: Overall accuracy % of single / multi-temporal, single / multi-polarization images obtained for training data. D1 to D6 are the dates of the COSMO-SkyMed images. 3D1 to 3D4 are the groups of 3 images starting from date D1 up to D4.

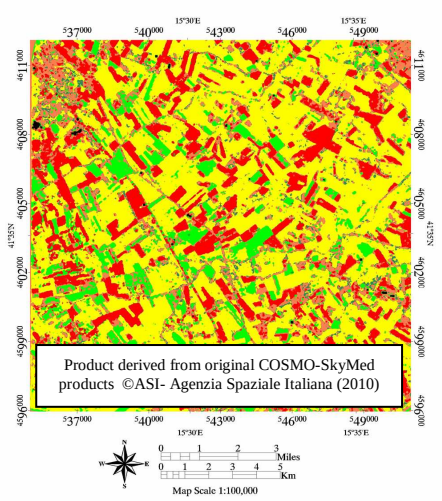


Figure 3: Land cover image obtained from HH+HV, multi-temporal COSMO-SkyMed images (acquisition dates 2010-05-29, 2010-07-08, 2010-07-24). Wheat, sugar beet, and tomato fields are in yellow, green and red, respectively.

4. Concluding remarks

This paper reported on the classification activities carried out during the first year of the COSMOLAND project. A multi-temporal series of X-band COSMO-SkyMed SAR data acquired over the Foggia agricultural site was used to investigate the potential of SAR data fro crop classification. Results showed that classification accuracies improve of 10%-20% by using multi- with respect to single-date X-band SAR data. Future work will be dedicated to extend the analysis to a larger set of test fields over the Foggia site, to longer time series of COSMO data and to the other agricultural sites included in the COSMOLAND project.

References

- McNairn H., Brisco B. (2004), The Application of C-Band Polarimetric SAR for Agriculture: a Review, *Canadian Jou. of Remote Sensing*, 30 (3): 525-542.
- Skriver, H., F. Mattia, G. Satalino, A. Balenzano, V.R.N. Pauwels, N.E.C. Verhoest, and M. Davidson (2011), Crop classification using short-revisit multitemporal SAR data, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (J-STARS)*, 4 (2), 423- 431.
- ASI ref. (2010), http://www.cosmo-skymed.it/it/index.htm

Multidimensional analysis of data from Bari Harbour: a GIS based tool for the characterization and management of bottom sediments.

P. Dellino[1], D. Mele[1], M. Mega[2], E. Pagnotta[2], F. De Giosa[3], G. Taccardi[3], N. Ungaro[4], G. Costantino[4].

[1]Università di Bari – Dipartimento Geomineralogico, Via E. Orabona 4 - 70125 Bari - Italy
[2]Autorità Portuale del Levante, P.le C. Colombo, 1 – 70121 Bari – Italy
[3]Coastal Consulting & Exploration, Via Aulisio, 59/61 - 70124 Bari - Italy
[4]ARPA Puglia, C.so Trieste 27 – 70126 Bari – Italy.

Abstract
Mediterranean harbours are today experiencing criticalities on the management of bottom sediment. Bari harbour among them is one of the most important in southern Italy, in terms of both commercial and touristic activities. Surveys dealing with the geophysical and sedimentological characterization of harbour sediments were performed during the period 2009-2011. The aim was: 1) a volume estimation of the sediments; 2) a granulometric characterization for the classification of sediment quality. The bathymetric and geophysical survey allowed a detailed estimation of the total sediment volume. The grain–size characterization consisted in the analysis of sediment samples collected at different depths in the harbour area. A combined elaboration of results lead to a multidimensional representation of the physical characteristic of sediments by means of a GIS platform.

Keywords: harbour sediments, geophysics, granulometry, harbour management.

1. Introduction

Bari harbour is located along the Southern Adriatic sea and is a multipurpose harbour (both commercial and passenger traffic). It is among the main Italian harbours, considering that it deals with a yearly commercial displacement of about 5 million tons (mainly dry generic goods) and about 2 million passenger transits (of which 600000 cruisers). The Harbour Authority of Levante, constituted according to the Italian law 84/1994, is in charge of port management. The hydrodynamic characteristics of the harbour and the geologic nature of its bottom substrate lead to a sediment circulation that provokes thickening of sediment near the entrance and docks. These sediment accumulations do not allow an optimum exploitation of the harbour operational depths. For this reason, maintenance dredging is necessary. The new legislative framework requires complex procedures for the obtainment of environmental permits, which dramatically slows down harbour's maintenance. In order to organize maintenance activities, a detailed knowledge of the harbour bathymetry, sediment thickness and grain size is needed. It is to remember that sediment disposal is regulated by severe environmental laws, especially for the pelitic fraction (< 0.0064 mm) (ICRAM-APAT, 2007).

2. Materials and Methods

A stratigraphic and bathymetric survey, followed by a coring campaign and sediment sampling was performed in the Bari harbour in the period 2009-2011. The stratigraphic and bathymetric survey, completed in 2009, was carried out for determining the status of the seafloor. The survey was certified according to the IHO Special Order S-44. The navigation and geophysical data acquisition system consists of a central computer equipped with two specific softwares (Thales PDS 2000, Communication Technology SwanPro), both interfaced with the positioning system, the sound velocity profiler, the high-resolution multi beam echosounder transducer and the single-channel sub-bottom profiler. Raw bathymetric data were processed by CARIS HIPS 7.0 sw, which allows the creation of a weighted grid surface (BASE, Bathymetry Associated with Statistical Error), reduced to the mean sea level as vertical datum reference. The stratigraphic survey was executed using the high resolution seismic reflection methodology.

The probing campaign with sediment sampling was carried out in the period January-February 2011. Cores of 10 cm radius were extracted in the sectors where, by the data of the previous bathymetric survey, the depth of the sea bottom resulted lower than the operational depth of the commercial and touristic traffic. The cores were of a length between about 1 and 2.5 m. Sediment samples were extracted each 50 cm. Most of the cores refer to the inner harbour perimeter and docks. An area near the harbour entrance was also cored. On the sediment samples, grain-size analyses were carried out. The grain-size distributions were represented as relative frequency distribution of percent weight and cumulative distributions.

3. Results

With regard the stratigraphic survey, the collected data consist of n. 87 seismic profiles (fig. 1) showing the geometry generated by the major acoustic reflectors and related to the interfaces between different sedimentary layers. Seismic data were processed by TEI sw in order to reconstruct the thickness of the sediments layer deposited on the bedrock through the picking operation. The modelling of the surfaces of the seabed and limestone bedrock respectively, allows to estimate the amount of loose sediment to be dredged in 120,000 m^3, in relation to established minimum safety depth for movement and berthing port areas.

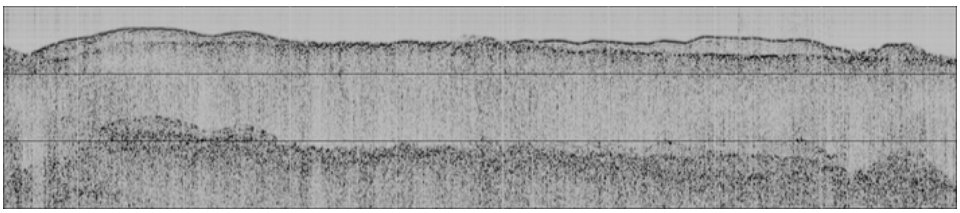


Figure 1: A seismic profile. The position of seismic reflectors allows locating the limit between the calcareous rock substrate with the overlying sediments.

The mosaic of the seismic profiles, performed by using GIS software platforms, allowed the 3d reconstruction of both the geometry of the rock substrate (fig 3a) and the sediment thickness (fig. 3b), allowing to highlight the main thickenings and accumulation of sediments near the harbour entrance and docks.

The grain-size spectrum from the sediment samples analysis covers a range between 2mm and 0.002mm, and is represented by means of the ϕ metric, where ϕ is = $-\log_2 d$, and d is particle diameter in mm. From the cumulative distribution the median size, Mdϕ (50^{th} percentile of the cumulative distribution) and sorting, σϕ (16^{th}-86^{th} percentile/2), which represent, respectively, a graphic approximation of the central tendency and of the dispersion of the distribution was calculated (fig.2).

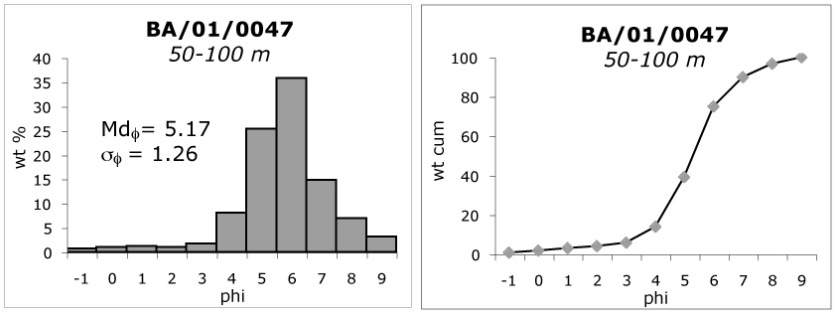


Figure 2: A sediments sample from the Bari harbour: an example of grain-size distribution histogram and cumulative distribution.

4. Concluding remarks

By combining data from the bathymetric investigation and from the granulometric characterization of sediments, it is possible to evaluate both the total volume of sediments inside the harbour and also the amount that needs to be dredged for the harbour maintenance. Furthermore, it is possible to highlight the relationship between sediment thickness and grain size, as shown on figure 3c. Data show a broad variability of grain size, both as a function of depth and location inside the harbour. These data are to be interpreted with reference both to the net sediment supply as due to the marine currents and also as a function of sediment recirculation, inside the harbour, as due to the ships movement. Sediment recirculation is favoured in the front of docks and much attenuated on the docks rear. In conclusion, starting from these data and by means of further investigation, it will be possible to implement sediment circulation models in the various sector of Bari harbour, and the relative sedimentation rate, with the aim of better designing the dimension and effectiveness of maintenance dredging according to the available rules and guide-lines (AA.VV., 1999; ICRAM-APAT, 2007). These results represent a good base for the purpose of future integrated management of the investigated harbour.

References

AA.VV (1999). Methodological Guide for Monitoring and Management of Environmental Aspects in Ports Areas. *ECO-information in European ports- Contract WA-97-SC.1132*, Volume N° 1: 254 pp.

ICRAM-APAT (2007). Manuale per la movimentazione dei sedimenti marini. *Ministero dell'Ambiente e della Tutela del Territorio e del Mare*, 77 pp.

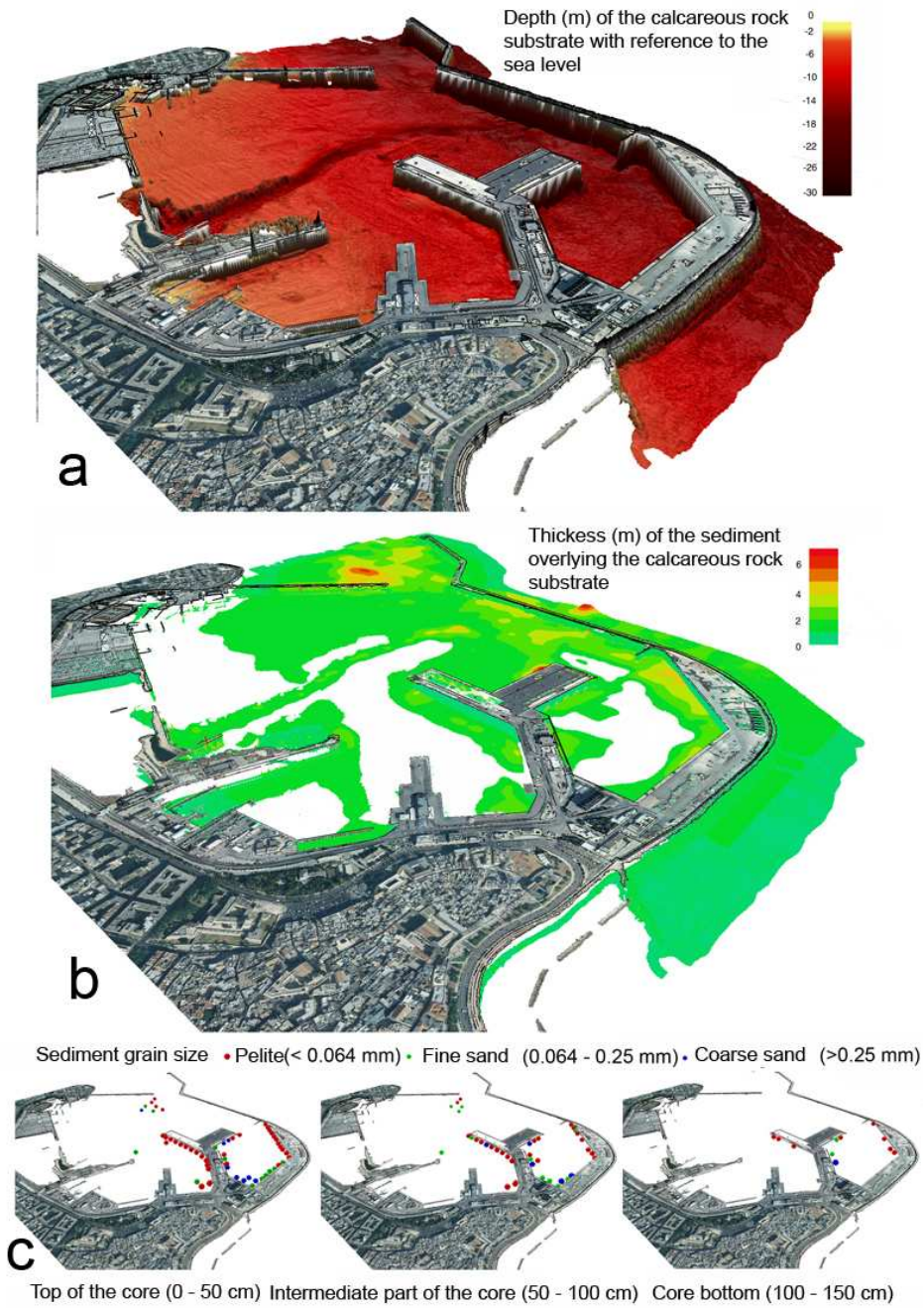


Figure 3: a = depth of the calcareous rock substrate. b = sediment thickness. c= sediment grain size.

Multivariate statistical analyses for the source apportionment of groundwater pollutants in Apulian agricultural sites

Pierina Ielpo, Daniela Cassano, Antonio Lopez, Pasquale Abbruzzese De Napoli, Giuseppe Pappagallo, Vito Felice Uricchio
Water Research Institute – National Research Council, via F. De Blasio, 5 – 70132 Bari, Italy; phone + 39 080 5820512; fax +39 080 5313365
piera.ielpo@ba.irsa.cnr.it

Abstract: Multivariate statistical techniques, such as Principal Component Analysis, Absolute Principal Component Scores, Cluster Analysis and Discriminant Function Analysis were applied to data set (pH, Electrical Conductivity, Total Dissolved Solids (TDS), Dissolved Oxygen (O2), Chemical Oxygen Demand (COD), the major ions (i.e. Na^+, Ca^{2+}, Mg^{2+}, K^+, Cl^-, NO_3^-, SO_4^{2-} and HCO_3^-), vital organism at 22 °C and 36 °C) of ground waters collected in 473 sites of the Apulia region during the "Expansion of regional agro-meteorological network" project. Multivariate statistical techniques allowed to identify for each province sites with different characteristics as respect to similar characteristics ones. Moreover Absolute Principal Component Scores allowed to identify generally three pollutant sources.

Keywords: ground water, water pollutants, source apportionment, statistical analyses

1. Introduction

During the years 2004-2007 the Agricultural and Food Authority of Apulia Region has implemented the project "Expansion of regional agro-meteorological network" in order to assess, monitor and manage the regional groundwater quality. The wells monitored during this activity amounted to 473 and the water samples analyzed were 998.
This resulted in a huge and complex data matrix comprised of a large number of physical-chemical parameters, which are often difficult to interpret and draw meaningful conclusions. Further, for effective pollution control and water resource management, it is required to identify the pollution sources and their quantitative contributions. The application of different multivariate statistical techniques such as cluster analysis (CA), principal component analysis (PCA), source apportionment by multiple linear regression on absolute principal component scores (APCS) for interpretation of the complex databases offers a better understanding of water quality in the study region. Moreover Discriminant Function Analysis (DA) was used in order to identify the characteristics of the all sites investigated in the Apulia region.

2. Materials and Methods

Groundwater samplings were performed under dynamic conditions, after flushing a large amounts of water for about 30 minutes. Samples were collected in polyethylene tanks with cap and under cap, filled to the brim in order to prevent the transfer of the analytes in the headspace and their loss at the opening of the tanks. After collection, samples were stored in cooled bags and transported to the laboratory as soon as possible.
The samples were analyzed for pH, Electrical Conductivity (Electr. Cond.), Total Dissolved Solids (TDS), Dissolved Oxygen (O2), Chemical Oxygen Demand (COD), the major ions (ie. Na^+, Ca^{2+}, Mg^{2+}, K^+, Cl^-, NO_3^-, SO_4^{2-} and HCO_3^-), vital organism at 22 °C and 36 °C, according to the official guideline proposed by the Ministero delle Politiche Agricole (the national agriculture authority) in a specific law (Decreto Ministeriale del 23 Marzo 2000 "*Metodi ufficiali di analisi delle acque per uso agricolo e zootecnico*"). Each parameter was analyzed in three replicates. In table 1 the number of monitored wells and collected samples for each Apulian province have shown.

Province	Wells	Samples collected
BARI	96	260
BRINDISI	89	102
FOGGIA	85	219
LECCE	84	165
TARANTO	119	252

Table 1: Groundwater quality monitoring

3. Results

DFA applied to all data set allowed to individuate the variables with bigger discriminatory power. The results are shown in table 2: among variables those with bigger discriminatory power are highlighted in bold.
PCA, CA, APCS methods were firstly applied to the samples collected in each Apulian province separately. Form results obtained by PCA and CA was evident that for each province some sampling sites investigated showed dissimilarities, mostly due to the location of the site (close to the sea, close to not purified water channels), the land use and management techniques (fertilizing and nourishing techniques) and groundwater overuse of the investigated sites. For all these reasons several natural and anthropogenic sources affect the groundwater quality of the investigated sites. As example some results of PCA and APCS for Taranto province are shown in figure 1 and 2.
Considering the score plot (figure 1) in the plane of the first and second Principal Component it is possible to note some scattered samples, highlighted in rectangular lines. Moreover considering the loading plot (not shown here) the samples (sites) enclosed in the red lines (figure 1) show high loading values for vital organism at 22 °C and 36°C; samples enclosed in the green line show high loading values for TDS, Electr. Cond., Cl^-, Na^+, Mg^{2+}; samples enclosed in the brown line show high loading values for COD, SO_4^{2-}, Ca^{2+} and those in magenta line show high values for K^+.

Variables	Wilks Λ	Partial □Λ
pH	0.3488	0.6743
Elect Cond	0.2429	0.9684
TDS	**0.2376**	0.9900
O_2	0.2379	0.9886
Na^+	**0.2355**	0.9987
Ca^{2+}	**0.2359**	0.9971
Mg^{2+}	0.2438	0.9649
K^+	0.2529	0.9301
COD	**0.2365**	0.9945
Cl^-	**0.2353**	0.9995
NO_3^-	0.2388	0.9851
SO_4^{2-}	0.2382	0.9876
HCO_3^-	0.2416	0.9737
Vit. Org. 22°C	**0.2362**	0.9956
Vit. Org. 36°C	**0.2356**	0.9983

Table 2: Variables selected by Wilks lambda

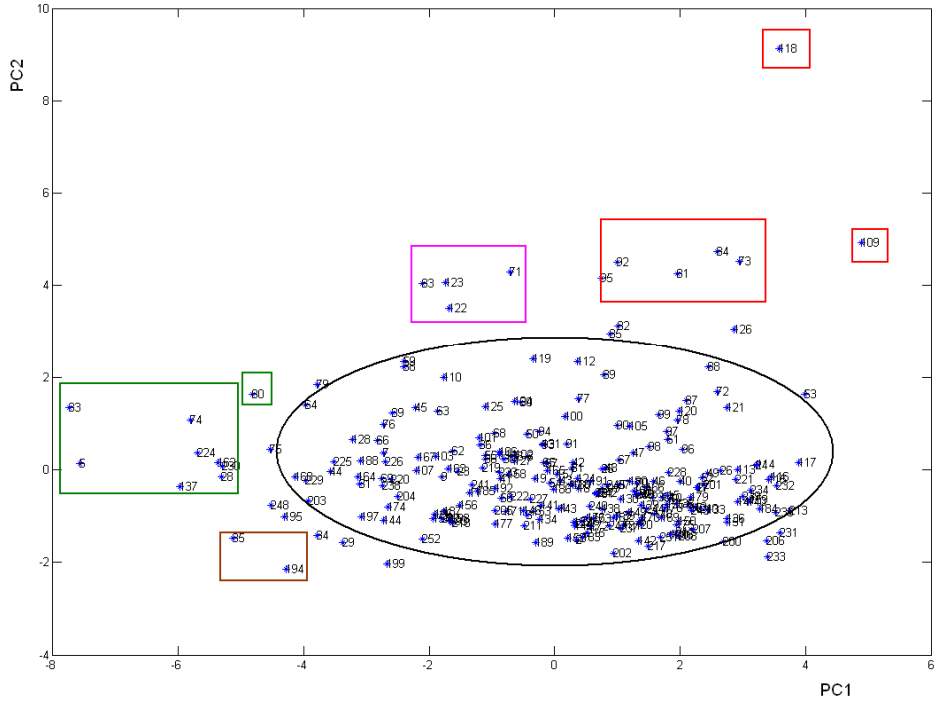


Figure 1: Score plot for Tarano province data matrix

The sites in red line are located close to a channel collecting waters from municipal purifier plant; those in green line are located close to the coast and for those it's possible

suppose an intrusion of marine water. The sites enclosed in brown line are located in an area with high agricultural impact: this means high use of fertilizers and nutrients.
The CA results support the PCA ones.

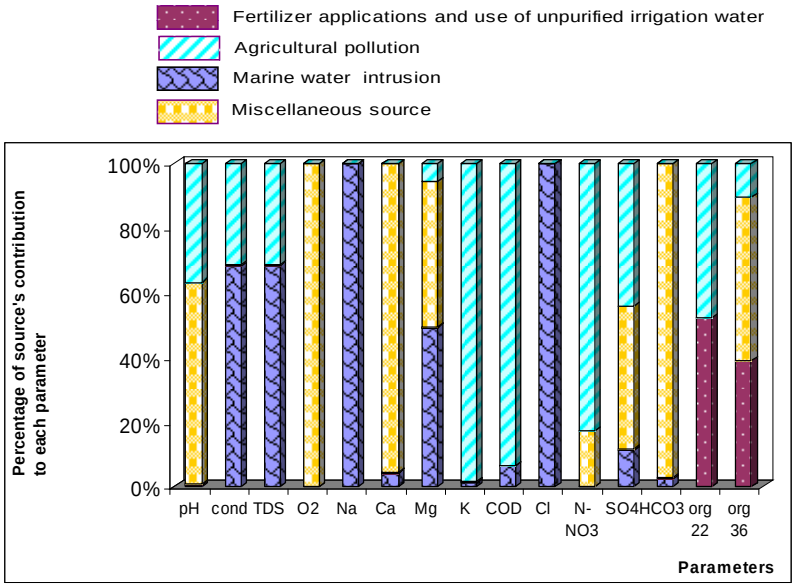


Figure 2: Percentage of source's contribution to each parameters for Taranto province data matrix

In order to individuate the pollutant sources the APCS method was applied to the data matrix of physical-chemical parameters collected. By APCS method it's been possible generally to identify three pollutant sources. About Taranto province data matrix (see figure 2) the pollution sources identified were: a source due to fertilizer applications and use of unpurified irrigation water; a pollution source due to agricultural techniques, marine water intrusion in the site one and a source mostly due to the calcareous characteristics of the soil in that area.

4. Concluding remarks

Multivariate statistical methods represent a valid tool to understand complex nature of groundwater quality issues, determine priorities in the use of ground waters as irrigation water and suggest interactions between land use and irrigation water quality. The results obtained by multivariate statistical methods can be used to suggest to stakeholders, for example, a mitigation in the groundwater overuse of some wells mostly in dry seasons and to require orderly quality tests of the channel waters when they are used for crop irrigation.

References

Bengraine K., Marhaba T. F. (2003) Using principal component analysis to monitor spatial and temporal changes in water quality, *Journal of Hazardous Materials*, 100, 179-195.

Structural changes in seismic activity before large earthquakes

Marta Gallucci, Alessandra Petrucci
Dipartimento di Statistica "G. Parenti", Università di Firenze
viale Morgagni 59, 50134 Firenze (Italy) - gallucci@ds.unifi.it

Abstract: We try to verify whether significant changes in the seismic activity are identifiable prior to a main shock, by means of statistical tests for structural changes applied to earthquakes models. A panel of models is selected, ranging from zero-inflated Poisson model to temporal and spatio-temporal point processes.

Keywords: seismicity patterns, point processes, structural changes

1 Introduction

As Kanamori (1981) pointed out, "various seismicity patterns before major earthquakes have been reported in the literature". Indeed, there is not uniform empiric evidence about the observed seismic activity in proximity of a relevant earthquake. Leaving aside differing definitions, some kinds of pattern are diffusely identified, sometimes following one another in the same seismic sequence: foreshocks (a large number of small events clustered in the main shock area), quiescence (reduced seismic activity before a large event), precursory swarms (distinct clusters of small earthquakes) and doughnut patterns (a quiet focal area surrounded by a region characterized by intense activity). However, different observations have in common the presence of a change in the seismic activity before a major event. Our aim is to detect these changes by means of statistical tests for structural breaks.

2 Materials and Methods

We choose to focus on the area surrounding the city of L'Aquila to verify the assertion of Papadopoulos et al.(2010). In their analysis of the seismic sequence prior to L'Aquila (Italy) earthquake (Mw 6.3) of 6th April 2009, the authors claim to have observed a change in the seismicity rate (daily number of events).

Data on earthquake events for Italy are publicly available on the website of the Istituto Nazionale di Geofisica e Vulcanologia (INGV). The area is identified as the square with side length of 100 km centered in the conventional coordinates of the city of L'Aquila (Lat. 42.35, Lon. 13.40), corresponding to the Forte Spagnolo.

The simplest way to describe a sequence of earthquake events in statistical terms is to consider the number of occurrences in periods of a certain length (days, weeks, months), which can be described in terms of a Poisson distribution. However, if the selected length is short enough, then a considerable number of periods with no events is present. To model this characteristic, we will refer to a zero-inflated Poisson (ZIP) model, as proposed by Guillas et al. (2010) in their analysis of the relationship between the ENSO and EPR seismicity. The number of events in the i-th period, Y_i, is described as follows:

$$\begin{cases} Y_i \sim 0 & \text{with probability } p \\ Y_i \sim Poisson(\mu) & \text{with probability } 1-p \end{cases} \tag{1}$$

To consider the presence of autocorrelation, neglected by the basic ZIP model, we will also perform time-series analysis, firstly introducing an auto-regressive component.
However, earthquake occurrences are more frequently described by means of point processes. Among this class of models, the most known is the Epidemic-Type Aftershock Sequence (ETAS) proposed by Ogata (1988). The first version of the model did not consider spatial coordinates, which were introduced in successive works (e.g. Ogata 1998). In this model, every single event is susceptible to produce an aftershock sequence. The occurrence rate of events, that is the conditional intensity λ, is therefore given by background seismicity μ and aftershocks, which are a function of magnitude (M) and of spatio-temporal $(x; y; t)$ coordinates of the triggering events:

$$\lambda(t, x, y) = \mu(x, y) + \sum_{j:t_j<t} v(t - t_j) \times g(x - x_j; y - y_j; M_j - M_c) \tag{2}$$

where M_c is the cutoff magnitude. Various kinds of $g(\cdot)$ functions are proposed.

Several tests for structural breaks are available in literature (e.g. by Chow, Quandt and Brown, Evans and Rubin). As noted by Hansen (1990), test procedures based on repeated estimation of the model are demanding, and not practically applicable, in case of complex models, which require relevant computational efforts. Therefore the author proposes a Lagrange multiplier (LM) test which only requires to estimate the model under the null hypothesis of no variations in the parameters.

3 Results

The analysis of the data leads to identify two periods when the seismic activity is appreciably intensified: the first between the end of 1997 and the first half of 1998, when the area was partially affected by the earthquake sequence in Umbria and Marche, and the second since April 6th main shock in L'Aquila. These two events

clearly represent a departure from the normal seismicity in the area, and shall be excluded from the analysis to avoid conditioning the results. Therefore, we consider the period from June 1st, 1998 and March 31th, 2009.

The first analysis accomplished with the ZIP model confirm the importance of considering the temporal correlation, but they are in contrast with the claim for a change on October 28th 2008 (see Table 1), as stated by Papadopoulos et al., while a change at the end of March 2009 seems to be more reasonable (see Table 2).

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
μ regression				
Intercept	-0.136	0.036	-3.778	0.000
AR(1)	0.106	0.012	8.810	$< 2 \cdot 10^{-16}$
p regression				
Intercept	-0.347	0.072	-4.855	$1.2 \cdot 10^{-6}$
Break date: 2008/10/28	0.136	0.118	1.148	0.251
	$AIC = 8244$	$p = 0.41402$ (SE 0.02457)		

Table 1: ZIP model (1) – no significant break is detectable on October 28th 2008.

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
μ regression				
Intercept	-0.132	0.036	-3.689	0.000
AR(1)	0.088	0.013	6.695	$2.2 \cdot 10^{-11}$
p regression				
Intercept	-0.363	0.073	-4.998	$5.8 \cdot 10^{-7}$
Break date: 2009/03/25	1.133	0.266	4.264	$2.0 \cdot 10^{-5}$
	$AIC = 8220$	$p = 0.41028$ (SE 0.02525)		

Table 2: ZIP model (2) – a significant break occurs on March 25th 2009.

ETAS model analysis is in a preliminary stage, but similar results seem to arise: Figure 1 shows a change in the residual process at the end of the considered period.

4 Concluding remarks

Due to different seismicity patterns empirically observed, the result of the analysis may vary with respect to the location and the extension of the considered area.

With respect to the period and the location we consider, and albeit further investigation is necessary, we shall doubt about a significant change occurred at the end of October 2008, but we can expect it to be identifiable in late March 2009.

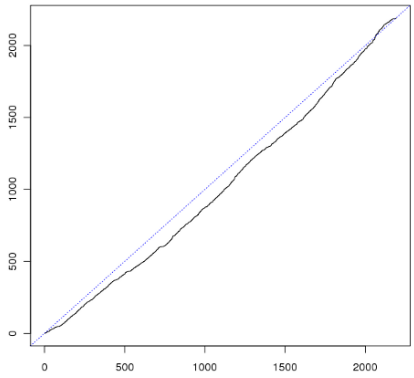


Figure 1: Residuals of the ETAS model

References

Brown R.L., Durbin J., Evans J.M. (1975) Techniques for testing the constancy of regression relationships over time, *Journal of the Royal Statistical Society*, B37, 149-163

Chow G.C. (1960) Tests of Equality Between Sets of Coefficients in Two Linear Regressions, *Econometrica*, 28, 591-605

Guillas S., Day S. J., McGuire B. (2010) Statistical analysis of the El NioSouthern Oscillation and sea-floor seismicity in the eastern tropical Pacific, *Philosophical Transactions of the Royal Society A*, 368, 2481-2500

Hansen B.E. (1990) Lagrange Multiplier Tests for Parameter Instability in Non-Linear Models, in: *Sixth World Congress of the Econometric Society*

Kanamori H. (1981) The nature of seismicity patterns before large earthquakes, in: Simpson D.W., Richards P.G., Earthquake prediction: an international review, *American Geophysical Union*

Ogata Y. (1988) Statistical models for earthquake occurrences and residual analysis for point processes, *Journal of the American Statistical Association*, 83, 9-27

Ogata Y. (1998) Space-time point-process models for earthquake occurrences, *Annals of the Institute of Statistical Mathematics*, 50, 379-402

Papadopoulos G.A., Charalampakis M., Fokaefs A., Minadakis G. (2010) Strong foreshock signal preceding the L'Aquila (Italy) earthquake (Mw 6.3) of 6 April 2009, *Natural Hazards and Earth System Sciences*, 10, 19-24

Quandt R.E. (1960) Tests of the Hypothesis that a Linear Regression System Obeys Two Separate Regimes, *Journal of the American Statistical Association*, 55, 324-330

Using environmental metrics to describe the spatial and temporal evolution of landscape structure and soil hydrology and fertility

Juan Antonio Pascual Aguilar
Centro para el Conocimiento del Paisaje, Rocha del Cine-41, Matet, 12415 Castellón, Spain, e-mail: juanantonio.pascual@imdea.org

Juana Sanz García
Irene de Bustamante Gutierrez
Malaak Kallache
IMDEA-Agua, Geomatics Unit, Calle Punto Net 4-2ª planta, Edificio ZYE, Parque Científico Tecnológico de la Universidad de Alcalá, Alcalá de Henares, 28805 Madrid, Spain, e-mail: juana.sanz@imdea.org, irene.bustamante@imdea.org, malaak.kallache@imdea.org

Abstract: In this work a methodology using Geographical Information Systems was developed and applied to a temporal series of land cover layers (for the years 1956, 1978, 1991 and 2010) in the municipality of Vall d'Uxó, Eastern Spain. Four types of metrics were implemented (1) spatial representation of the degree of artificialisation, (2) patchiness and fragmentation, (3) fertility dynamics of soils according to their land capability, and (4) soils imperviousness and loss of water retention capacity.
Results showed that the set of metrics can efficiently represent spatial and temporal dynamics. Furthermore, a link can be distinguished between trends in the degree of artificialisation, landscape structure and soil fertility and water retention properties for the region analysed.

Keywords: Land use-cover change, anthropogenic soil sealing, land degradation, spatial landscape metrics

1. Introduction

Land use-cover change is an ongoing process in both time and space. In recent years, land cover dynamism has accelerated in urban area hinterlands, being notorious in the Mediterranean region. It has been observed that the major mechanisms of change in western Mediterranean areas are intensification and transformation, which can be integrated in the general trend of artificialisation (Pascual Aguilar, 2002). While intensification mainly concerns the change from traditional rain-fed agriculture to cash crop irrigation practices, transformation is related with the substitution of one type of land use (and subsequent cover) by another, as in the transition from cultivated fields to buildings and roads.
According to the European Union perspectives, a transformation trend known as anthropogenic soil sealing is one of the most worrying aspects of soil degradation, along with the loss of useful biota, affecting desertification in dry environments where rainfall

is scarce. One of the major impacts of soil sealing is the loss of fertility and the alteration of the water regime due to imperviousness of the top soil layers.
Approaches to help understand pattern dynamics of land use-cover changes has been developed. Less studied are the interactions between trends in such spatial metrics and the environmental effects of land use-cover dynamics on soils and their water regimes.
The general aim of this research is thus the development of a descriptive framework based on landscape and environmental metrics to assess land use-cover spatial and temporal dynamics. Specific objectives are the application of spatial environmental metrics to analyze historical trends in (1) anthropogenic soil sealing, (2) in landscape structure changes, and (3) in soil productivity and soil water dynamics.
The analysis has been applied to the municipality of Vall d'Uxó in the province of Castellón, Eastern Spain. It is located in a transition area between Mediterranean coastal plains and pre-littoral mountain ranges. In recent decades, the region has undergone an intense dynamism dominated by the transition from traditional agricultural systems to highly technified irrigated cash crops and artificial surfaces (Pascual Aguilar, 2002), both processes identified in other regions and described respectively as intensification and conversion (Lambin, 1997).

2. Materials and Methods

Several layers of information were built up using conventional Geographical Information Systems software (ArcGis 9.3). Initial maps consisted of (1) detailed (scale 1:10000) land cover layers for the years 1956, 1978, 1991 and 2010, and (2) the construction of a soil map according to FAO nomenclature from published reports (Rubio et al. 1995).
Initial data were further processed to obtain layers with artificial surface urban and infrastructures classes for the respective land cover years and, from the soil maps one layer with soil agricultural capabilities following existing well established methodologies (Antolín, 1998) and a second one with soil water retention properties were extracted from the information from samples provided in soil reports (Rubio et al. 1995).
Four different types of metrics have been developed. First, landscape structure metrics were used. A cartographic value was developed, the Synthetic Index of Landscape Artificialisation (SILA), which expresses trends of artificial covers per unit area (a representative square of 100 x 100 m). The index includes under the same landscape class both agricultural intensification and paved and concrete surfaces and is the result of calculating the percentage of this class for each representative square area.
Second, based on existing metrics (e.g. Cushman et al. 2008), landscape structure was analyzed to determine the degree of fragmentation and patchiness with the specific landscape class of paved and built up surfaces, with major impact on soils and their water regime. The metrics applied to each year were the Number of Patches (NP), the Maximum Patch Size (MxPS in hectares), the Patch Average Size (MePS in ha), and the Patch Size Variance (PSV).
Third, specific metrics were developed to obtain insight about the potential environmental impact on soils fertility, understood as their capability to produce food.
Fourth, specific metrics have been also calculated to describe the impact of soil sealing on the water holding properties of soils. Metrics developed for land capability and water

retention are: Total Surface by Land Capability Type, TSLCT (in ha); Water Retention Capacity, WRC (in m^3); Cumulative Loss of Land Capability for a given year, CLLC (in ha); Cumulative Loss of WRC for a given year, CLWRC (in m^3);Ratio between WRC and TSLCT for a given year, WRC/TSLCT, and ratio between remaining water holding capacity (m^3) and remaining total land capability land (ha) for a given year, RWRC/RTSCU.

3. Results

In 54 years a landscape character change of considerable dimensions has taken place. The SILA index graphic expression (Figure 1) shows a constant increase in time and space with almost 50% of the reference squares above 50% of change due to agricultural intensification and artificial surfaces.

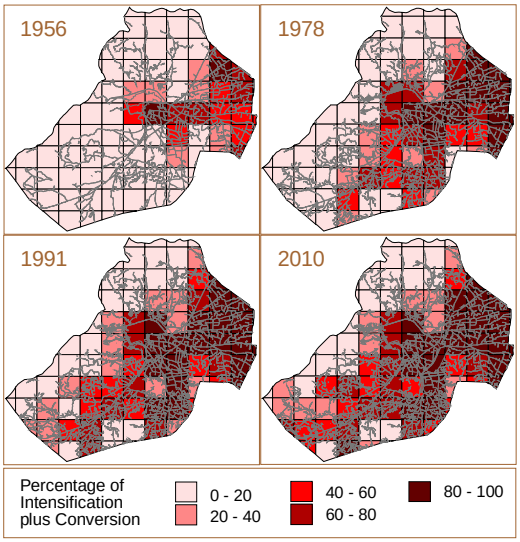


Figure 1: Cartographic representation of the Spatial Synthetic Index of Landscape Artificialisation

Landscape structure is analyzed by a set of four metrics (Number of Patches, NP; Mean Patch Size, MePS; Maximum Patch Size, MxPS; Patch Class Size Variance, PSV) (Table 1). All four series are monotone. A Mann-Kendall trend test with exact distribution of the test statistics, which is suitable for short time series (Hamed, 2009), suggests a significant trend at the 95% level for monotone series and a series length of four, as given here. We therefore regard these trends as significant, and the strength of the trend is approximated by the slope of a straight line, which is fitted to the data. The values, which are given in Table 1, suggest strong trends with constant increasing patchiness with time and consequent reduction of the remaining metrics (MePS, MxPS, and PCSV). The increase in number of patches results in the physical fragmentation of the initial landscape units and consequently, MePS, MxPS and PSV get smaller because there is a trend to reduce differences between patch sizes. Also, environmental consequences of the above trends are reflected in the reduction (the soil sealing process) of soils covered by natural or cultivated vegetation.

Relationships between landscape structure due to artificial landscape classes and soil fertility and hydrological properties are established by a new group of environmental metrics (Table 2 and Figure 2). They are related to five types of land capability to produce biota (Very High, A; High, B; Moderate, C; Low, D and Very Low, E). Fragmentation and patchiness are produced by the increment of artificial surfaces that substitute former soil covers of natural or agricultural landscape classes, which area synthesized by land capability A, B, C, D, and E types and represented by CLLC metric (Table 2). Also the anthropogenic sealing will cover the soil top layer avoiding water processes and soil moisture dynamics (CLWRC). Apparently these metrics are different for different landscape classes. For A-C the formula "the better the soil, the lower the yearly loss" can be established. However, the two lowest biota producing classes, D and E, have very low loss rates in 1960. Trends in decline in soil fertility are evident with time. Land capability classes C, D and B have lost greater proportion of soil fertility and water holding efficiency. Due to this trend, the low biota producing class D undergoes a dramatic loss of soil fertility and water holding efficiency: in 2010 it has the second-highest losses. The largest losses occur for class C. All land capability classes experience stronger losses from 1990 on. Moreover, the RWRC/RTSCU ratio reflects a general trend in the decline of both indicators. However, the trend in the ratio is not strong, as gets apparent in Figure 2.

Metrics	**Year**				**Slope of Trend**
	1956	**1978**	**1991**	**2010**	
NP: Number of Patches	274.0	789.0	1641.0	1725.0	29.2
MePS: Mean Patch Size (ha)	24.7	8.6	4.1	3.9	-0.4
MxPS: Maximum Patch Size (ha)	1583.3	1203.1	1007.1	997.2	-11.2
PSV: Patch Class Size Variance	16794.6	3549.9	1386.7	1316.8	-286.0

Table 1: Synthetic landscape structure metrics

Land capability type	**Soil fertility and soil hydrology metrics**															
	TSLCT (ha)	**WRC (m3)**	**WRC/ TSLCT**	**CLLC (%)**				**CLWRC (%)**				**RWRC/RTSCU**			
	Reference situation			1956	1978	1991	2010	1956	1978	1991	2010	1956	1978	1991	2010
A	1441.8	24077678	1670.0	3.9	4.0	4.4	10.6	3.9	4.0	4.4	10.6	1670.0	1670.0	1670.0	1670.0
B	1083.7	1809312	1669.6	5.7	8.6	10.2	18.0	5.7	8.6	10.2	18.0	1669.6	1669.6	1669.6	1669.5
C	681.3	828819	1216.6	11.8	23.3	27.1	29.6	16.0	32.1	40.2	43.7	1158.1	1077.2	997.8	972.4
D	1408.7	726233	515.5	1.2	4.4	6.9	21.7	1.3	6.9	12.8	32.0	514.9	501.7	482.9	447.5
E	2209.0	1066269	482.7	1.4	2.5	3.6	4.6	1.4	2.4	3.4	4.4	482.9	483.2	483.6	483.4
Totals	6824.4	6838400	1002.1	3.6	6.3	7.9	14.0	5.2	8.7	11.0	17.9	985.9	976.0	967.6	956.9

TSLCT: Total Surface by Land Capability Type. WRC: Water Retention Capacity. CLLC: Cumulative Loss of Land Capability for a given year. CLWRC: Cumulative Loss of WRC for a given year. WRC/TSLCT: Ratio between WRC and TSLCT for a given year. RWRC/RTSCU: ratio between remaining total water holding capacity (m^3) and remaining total land capability land (ha) for a given year.

Table 2: Metrics related to soil fertility and soil hydrology

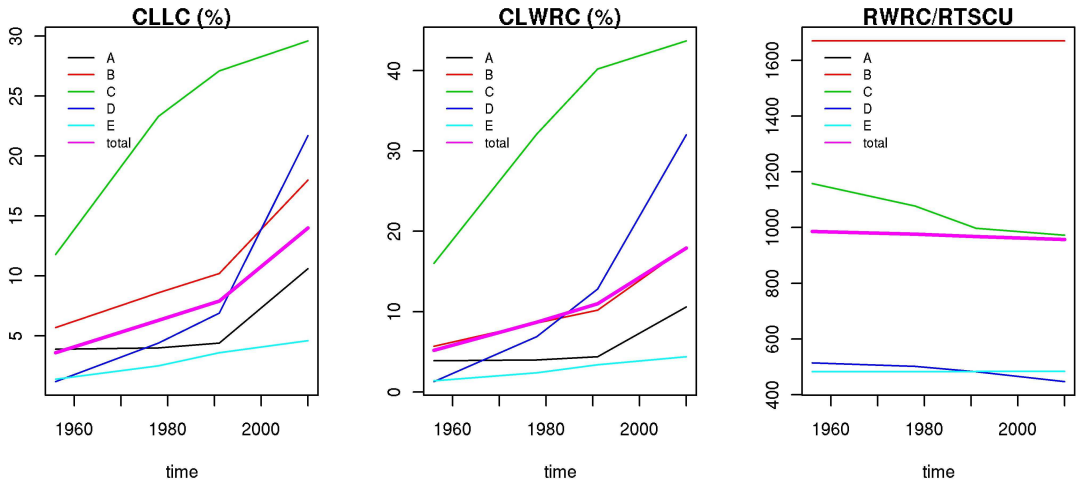


Figure 2: Metrics related to soil fertility and soil hydrology

4. Concluding remarks

Based on the process of landscape artificialisation due to land use-cover dynamics, the methodology developed a set of simple spatial metrics to relate potential impacts on soil fertility and hydrology. We found that the landscape structure for the region of Vall d'Uxó has become increasingly scattered over the last 50 years. Moreover, the link between conventional landscape pattern and structure metrics to new specific ones for land capability and water holding properties in soils describes the relation between land cover dynamics (in time and space) and their environmental interactions. We found different soil fertility and water holding capacity losses for different land types, which are distinguished according their potential to produce biota. However, an enhanced loss of all metrics within the last 20 years is identifiable for all land types.

References

Antolín, C. (1998) *El suelo como recurso natural en la Comunidad Valenciana*, Generalitat Valenciana, Valencia, Spain.

Cushman S. A., McGarigal K. & C. Neel M. (2008) Parsimony in landscape metrics: Strength, universality, and consistency, *Ecological Indicators*, 8, 691-703.

Hamed K.H (2009). Exact distribution of the Mann-Kendall trend test statistic for persistent data. *Journal of Hydrology* , 365:86-94.

Lambin E. F. (1997) Modelling and monitoring land-cover change processes in tropical regions, *Progress in Physical Geography*, 21, 375-393.

Pascual Aguilar J.A. (2002) Modelling the impact of land cover changes on the soil water regime, in: *Man and soil at the Third Millennium*, European Society for Soil Conservation, Valencia, Spain, Volume 1, 423-433.

Rubio Delgado J. L., Sánchez Díaz J. & Forteza Bonnin J. (1995) *Mapa de suelos de la Comunidad Valenciana. Sagunto (668)*, Generalitat Valencian, Valencia, Spain.

A comparison between hierarchical spatio-temporal models in presence of spatial homogeneous groups: the case of Ozone in the Emilia-Romagna Region[1]

Francesca Bruno, Lucia Paci
Dipartimento di Scienze Statistiche, Università di Bologna, lucia.paci2@unibo.it

Abstract: Hierarchical spatio-temporal models permit to estimate many sources of variability. In many environmental problems, different features characterizing spatial locations can be found. Differences in these classifications can show discrepancies either in mean levels or in the spatio-temporal dependence structure. When these characteristics are not included in the model structure, model performances and spatial predictions may lead to poor results. Here, we compare alternative enrichments of the hierarchical spatio-temporal model that consider the presence of groups. Our application concerns Ozone data in the Emilia-Romagna region in which the monitoring sites can be classified according to their relative position with respect to traffic emissions.

Keywords: spatio-temporal models, hierarchical models, groups of sites, ozone data.

1. Introduction

Hierarchical models, being very flexible, are suitable for dealing with differences both at the measurement and the process level (Wikle, 2003).
In the following, we expand the general framework describing hierarchical spatio-temporal models for studying geostatistical data by the inclusion of domain classifications with respect to certain differentiating features (Wang *et al*., 2009). When studying air pollution, for example, monitoring stations may be differently located with respect to traffic or household density. This peculiarity can be modeled in a number of different ways. Models that allow for differences between groups of sites have recently been proposed (Cocchi and Bruno, 2010). In environmental applications: for example, Paci (2010) proposed a hierarchical spatio-temporal model for pollutants where the group differences were captured by the intercept of the model (*i.e.* difference in pollution levels between the urban and rural locations). In Sahu *et al*. (2006) a hierarchical space-time model for $PM_{2.5}$ that includes two spatio-temporal processes was proposed, where the first captures the background effects, and the second adds extra variability for urban locations by using the relationship between the response variable and suitable covariates (the population density, in this case).
Here the inclusion of groups in spatio-temporal models is formalized in a more general way. We describe alternative proposals for including group differences in hierarchical Bayesian models. The assessment of the consequences for spatial prediction under this innovation will be also considered.

[1] Work supported by the project PRIN 2008: New developments in sampling theory and practice, Project number 2008CEFF37, Sector: Economics and Statistics, awarded by the Italian Government.

This paper is organized as follows: the next section describes the Ozone dataset; Section 3 sketches the main models that include spatial groups; the final section presents the main results and some concluding remarks.

2. The Ozone Dataset

Tropospheric ozone is one of the most important pollutants when studying air quality. Here, the dataset consists of Ozone daily measurements (in $\mu g/m^3$) collected from 31 monitoring stations across the Emilia–Romagna Region in 2001. Monitoring sites can be classified according to traffic pollution exposure (D.M.A. 16/05/1996); the two groups consist of 17 background monitoring sites (denoted by "G1") and 14 sites characterized by their vicinity to traffic emissions (denoted by "G2"). Monitoring sites belonging to G1 are expected to measure higher Ozone levels than sites belonging to G2. Some meteorological covariates are available for each site and each time. In particular, one of the most correlated with Ozone is the daily mixing height, that will be included as a covariate in the model.

3. Model specification

Let $\mathbf{Y}^* = \{\mathbf{Y}^*(\mathbf{u}, t); \mathbf{u} \in (\mathbf{u}_1, \cdots, \mathbf{u}_{n^*}), t \in (1, \cdots, T)\}$ denote the log-Ozone concentrations for the generic location and time $(\mathbf{u}, t)$. We consider 27 of the 31 sites for estimation and 4 sites for prediction assessment (2 for each group). Let define $\mathbf{Y}$ as the Tn-dimensional subset of the original dataset under these specifications.
Following the usual hierarchical spatio-temporal specification (Banerjee *et al*. 2004), let

$$\mathbf{Y} = \mathbf{Z} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{Z}$ is the Tn-dimensional spatio-temporal process and $\boldsymbol{\varepsilon}$ is a Gaussian noise process $N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_{Tn \times Tn})$, representing the spatio-temporal measurement error structure via homoscedastic and independent components. Conditionally on $\mathbf{Z}$ and σ_ε^2 the distribution of $\mathbf{Y}$ is:

$$\mathbf{Y}|\mathbf{Z}\,,\ \sigma_\varepsilon^2 \sim N(\mathbf{Z}\,, \sigma_\varepsilon^2 \mathbf{I}_{Tn \times Tn})$$

The second stage of the hierarchy can be defined as the combination of a large scale spatio-temporal process ($\mathbf{m}$), a spatial effect ($\mathbf{W}$) and a temporal effect ($\mathbf{V}$):

$$\mathbf{Z} = \mathbf{m} + \mathbf{1}_{T \times 1} \otimes \mathbf{W} + \mathbf{V} \otimes \mathbf{1}_{n \times 1} \tag{2}$$

The expression for the Tn-dimensional trend component ($\mathbf{m}$) is:

$$\mathbf{m} = \mathbf{X}\boldsymbol{\beta} \tag{3}$$

where $\boldsymbol{\beta} = (\beta_0\,, \beta_1)'$ and $\mathbf{X}$ is a $Tn \times 2$ covariates matrix with unit values in the first column and daily mixing heights in the second column. The expression in (2) provides additive temporal and spatial effects (multiplicative on the original scale). The temporal random effect $\mathbf{V} = \big(V(1)\,, \cdots, V(T)\big)'$ and the spatial random effect $\mathbf{W} = \big(W(\mathbf{u}_1), \cdots, W(\mathbf{u}_n)\big)'$ capture respectively any spatial and temporal dependence which remains unexplained by the model for the mean (3). The distribution of the random effect $\mathbf{V}$ can be expressed via the multivariate distribution

$$\mathbf{V} \sim N\big(\mathbf{0},\ \sigma_v^2\, \mathbf{A}(\phi)\big) \quad \text{where} \quad \big(\mathbf{A}(\phi)\big)_{ij} = exp\big(-\phi\, \|t_i - t_j\|\big) \tag{4}$$

and σ_v^2 is the scalar variance of the temporal component; $\mathbf{A}(\phi)$ is the $T \times T$ correlation matrix defined by the exponential function.
The spatial random effect $\mathbf{W}$ is modeled as a Gaussian process

$$\mathbf{W} \sim N\left(\mathbf{0},\ \sigma_w^2\ \mathbf{H}(\delta)\right) \quad \text{where} \quad \left(\mathbf{H}(\delta)\right)_{ij} = exp\left(-\delta \left\|\mathbf{u}_i - \mathbf{u}_j\right\|\right) \tag{5}$$

and σ_w^2 is the scalar variance of the spatial process; $\mathbf{H}(\delta)$ is the $n \times n$ spatial exponential correlation matrix.
The model hierarchy is completed by the specification of noninformative prior distributions for the hyperparameters.
In the following subsections we propose two different specifications of model (1) – (5) (from now on called "Model (A)") in order to take groups into account.

3.1 Modeling differences in the trend component

When the differences between the two groups are captured by the average level, the discrepancies are developed from model (3), the large-scale process can be rewritten as:

$$\mathbf{m} = \alpha \mathbf{d}_{\mathbf{u} \in G1,(t=1,\cdots,T)} + \mathbf{X}\boldsymbol{\beta} \tag{6}$$

In (6) α is a scalar type-specific intercept and $\mathbf{d}_{\mathbf{u} \in G1,(t=1,\cdots,T)}$ is a *Tn*-dimensional vector collecting the dummy variables that classify the spatial sites into groups. The β_0 parameter represents the intercept for the sites belonging to G2 and $\alpha + \beta_0$ represents the intercept for the other group. This model will be referred to as "Model (B)".

3.2 Modeling differences in the spatio-temporal covariance structure

When differences in the spatio-temporal dependence structure are included in the model, alternative $\sigma_w^2\ \mathbf{H}(\delta)$ might be considered in (5). Matrix $\sigma_w^2\ \mathbf{H}(\delta)$ is constituted by blocks, with group-specific spatial variance matrices in the diagonal after reordering sites according to the groups. The most complex model includes an out-of-diagonal between-group variance block matrix, $\sigma_{w(G1,G2)}^2 \mathbf{K}\left(\delta_{G1,G2}\right)$, that is characterized by group parameters:

$$\sigma_w^2\ \mathbf{H}(\delta) = \begin{bmatrix} \sigma_{w(G1)}^2\ \mathbf{H}(\delta_{G1}) & \sigma_{w(G1,G2)}^2\ \mathbf{K}\left(\delta_{G1,G2}\right) \\ \sigma_{w(G1,G2)}^2\ \mathbf{K}\left(\delta_{G1,G2}\right) & \sigma_{w(G2)}^2\ \mathbf{H}(\delta_{G2}) \end{bmatrix} \tag{7}$$

Specification (7) needs the estimation of a huge number of parameters. When interactions between locations belonging to different groups are ignored, $\sigma_{w(G1,G2)}^2 \mathbf{K}\left(\delta_{G1,G2}\right)$ is fixed at zero (in what follows "Model (C)").

3. Results

The comparison between models is performed both in terms of goodness of fit (via DIC) and in terms of predictive assessment (via Predictive Model Choice Criterion, PMCC, Sahu *et. al* 2006).Table 1 shows that Model (B) has the best performance. This highlights that the main differences between groups concern the mean levels and it is reasonable to assume a common correlation structure for both groups.

	Model (A) for G1 sites	Model (A) for G2 sites	Model (A) for all sites	Model (B)	Model (C)
DIC	6403	5737	11840	11830	11840
PMCC	187.50	231.03	391.32	377.13	400.99

Table 1: DIC and PMCC for all models considered

Figure 1 shows the predictions for a specific site and for all models. The predictive performances are similar for all models, the prediction credibility bands contain almost always the observed values.

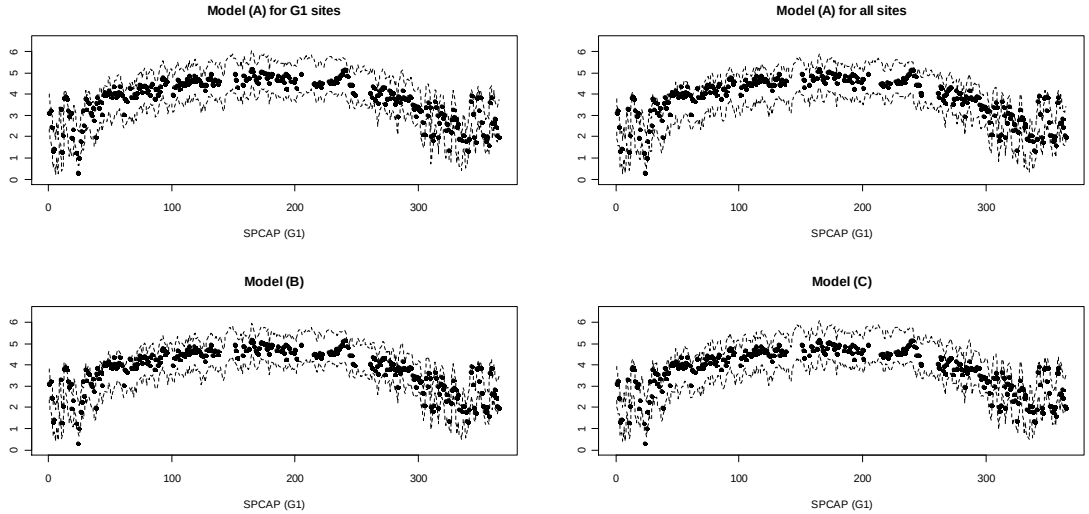


Figure 1: Predictions for a site belonging to G1 for 2001, estimated for all models

References

Banerjee S., Carlin B.P., Gelfand A.E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall, CRC Press.

Cocchi D., Bruno F. (2010) Considering groups in the statistical modeling of spatio-temporal data, *Statistica*, 4, *in press*.

D.M.A. (16/05/1996) Attivazione di un sistema di sorveglianza di inquinamento da Ozono. In Italian.

Paci L. (2010) Hierarchical Bayesian space-time model: the case of Ozone in Emilia Romagna, Thesis of master in statistics (in Italian).

Sahu S.K., Gelfand A.E., Holland D.M. (2006) Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 61-86.

Wang J., Christakos G., Hu M-G. (2009) Modeling spatial means of surfaces with stratified nonhomgeneity, *IEEE Transactions on Geoscience and Remote Sensing*, 47, 4167-4174.

Wikle C.K. (2003) Hierarchical models in environmental science, *International Statistical Review* 71, 181-199.

A multilevel multimember model for smoothing a disease map of lung cancer rates

Nicola Bartolomeo, Paolo Trerotoli, Gabriella Serio.
Department of Biomedical Science and Human Oncology, Chair of Medical Statistics, University of Bari. Bari. Italy.
E-mail of corresponding author: nicola.bartolomeo@uniba.it

Abstract: Aim of this study is to assess the effect of smoothing a hospitalization rates map, based on the assumption that they may be influenced by the neighboring municipalities, the health service organization (HSO) and environmental risk factors.
To smooth rates, two different Multilevel Multimembership Models were fitted: in the first the random effects were the municipality heterogeneity, the spatial dependence of the municipalities and the local HSO; in the second we replaced the local HSO effect by the environmental risk effect. The models were applied to show the spatial rates of hospitalization for lung cancer in Apulia in the year 2006.
Maps shaded with the rates obtained at the end of the smoothing procedure seem to express a geographic distribution pattern of higher or lower rates in specific areas of the region. The effect of smoothing was greater in municipalities with a more unstable Risk Adjusted Rate.

Keywords: Spatial analysis, Lung cancer, Smoothing, Multilevel Model

1. Introduction

Spatial analysis is often used to assess mortality or hospitalization rates but in such cases a problem of instability arises when they are calculated on small areas, owing to the small number of expected and observed cases (Olsen, 1996). Spatial smoothing could help to generate a correct interpretation of geographic variations of the risks of hospitalization or mortality (Carrington, 2007).
The primary aim of this study was to show, by spatial representation, how the hospitalization rates can be influenced both by the immediately neighboring municipalities and by the local health service management (ASL) to which the municipality belongs, as well as by environmental risk factors associated with the disease under study. As an example, the hospitalization rates for lung cancer recorded for the Apulia region were used.

2. Materials and Methods

To estimate the spatial effects with a multilevel model, the model must contain two components specifying the structure of random effects: a random effect or heterogeneity term, and a term representing the spatial contribution of neighborhood areas.
Because relative risks can be spatially autocorrelated, the multilevel model must be seen as a "Multiple Membership Model" (Goldstein, 2003; Goldstein, 1998), where each *municipality* belongs to a higher level unit that also contains the neighboring *municipalities*. The criterion used to establish the cluster level was the distance radius (25 km) within which all the *municipalities* are considered to belong to the same cluster.
Let's consider the i-th municipality with E_i expected cases obtained at the end of a procedure of Risk Adjustment by gender and age. The Multiple Membership Model is:

$$\log(\mu_i) = \log(E_i) + \alpha + x_i\beta + v_i + u_i \qquad (1)$$

where log (E_i) is treated as an offset, α is a constant, v_i represents the random effects due to the spatial dependency, u_i represent the effects of the heterogeneity among the *municipalities* and $x_i\beta = 0$ if there is no covariate.

Each *municipality i* is spatially dependent on one or more *municipalities j* belonging to the higher level geographic area ∂_i , each contributing with weight z_{ij} . The sum of the weights of *municipality i* is equal to one. Therefore, when drawing up the model each spatial effect v_i referred to *municipality i* must be taken as the sum of a set of independent random effects, so that:

$$v_i = \sum_{j \in \partial_i} z_{ij} v_j^* \qquad (2)$$

v_j^* can be seen as the effect of *municipality j* on the other *municipalities* and z_{ij} is its associated weight.

In our first hypothesis the hospitalization rate varies among *municipalities* also according to the different management of the diagnosis by the local health service units. For this reason, we added a further random effect w_i representing the ASL each *municipality* belongs to:

$$\log(\mu_i) = \log(E_i) + \alpha + \sum_{j \in \partial_i} z_{ij}\, v_j^* + u_i + w_i \qquad (3)$$

After building the matrix of random effects (Langford, 1999) Model A was estimated by (3) and the smoothed hospitalization rates for each *municipality* were calculated.

In equation (3), the parameters were estimated by the maximum likelihood technique.

In the second hypothesis the hospitalization rates vary among *municipalities* according to the degree of exposure to some risk factors. We identified 12 mutually esclusive areas of environmental risk, each centered around a *municipality* where industries with a high environmental impact are located, and extending for a radius of 10 km around it (Dominici, 2006). Then Model B was estimated by (3), with the random effect w_i that represents the risk area in which the *municipality* is located. Industrial poles with a high environmental impact are indicated on the proposed maps to explore their effect on the geographic distribution of the disease. The analysis was conducted by selecting, from the Hospital Discharge Forms (HDF) for Apulian residents for the year 2006, those patients admitted with ICD9-CM codes of primary diagnosis 162--. To fit the multilevel models we used the SAS software.

3. Results

In 2006, a total of 2,591 patients resident in Apulia were hospitalized with a primary ICD9-CM diagnosis in the category “Malignant tumors of the trachea, bronchi and lungs” (crude regional rate = 6.36 per 10,000 inhabitants). The parameters and estimated standard errors with Models A and B are shown in Table 1.

In Model A, the only significant parameter was the variance due to the municipality heterogeneity (p=0.0092). The spatially structured variability quota is lower: 18.27% (0.0091/0.0498), while the ASL value is equal to 1.00% (0.0005/0.0498).

In Model B the estimated random effect due to the environmental risk areas is not significant (p=0.2261), nor is the heterogeneity variance (p=0.0654), while the only significant parameter is the clustering variance (p=0.0183). The spatially structured variability quota is equal to 47.32% (0.0247/0.0522) and the environmental risk area variability is 0.57% (0.0003/0.0522).

	A - Model with spatial effect and ASL effect		B - Model with spatial effect and risk Area effect	
	Estimate	St. Error	Estimate	St. Error
Fixed part				
Intercept	-0.0369	0.0575	0.0826	0.0759
Random part				
$\sigma^2_{u\ \text{heterogeneity}}$	0.0402*	0.0154	0.0272	0.0147
$\sigma^2_{v\ \text{clustering}}$	0.2635	0.2498	0.7179*	0.3043
$\sigma^2_{w\ \text{ASL}}$	0.0218	0.0203		
$\sigma^2_{w\ \text{environmental risk area}}$			0.0064	0.0085
n_i	*29.07*		*29.07*	
m_i	*43.00*		*19.85*	
σ^2_v / n_i	0.0091		0.0247	
σ^2_w / m_i	0.0005		0.0003	
σ^2_{TOTALE}	0.0498		0.0522	
AIC	389.4		393.2	

* $p<0.05$

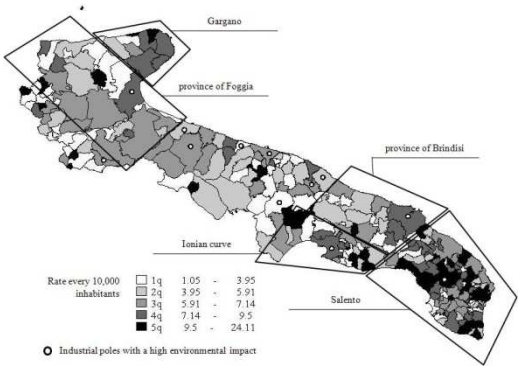
Table 1: Parameters and estimated standard errors in the rates smoothing models

Three maps were built: the first one using the rates obtained at the end of the Risk Adjustment procedure before smoothing and the second and third using the smoothed rates obtained after estimating Models A and B, respectively (Figure 1).

The map in figure 1a does not offer a clear visual picture of areas with higher or lower hospitalization rates for lung cancer. In figure 1b it can be seen that there is a tendency toward clustering of municipalities with a higher admission rate for lung cancer in the Salento, the southernmost part of the Ionian curve and the Gargano. In figure 1c the introduction of the random effect of the areas at environmental risk produces little variation in the appearance of the municipalities hospitalization rate level as compared to figure 1b. The Gargano area is differently highlighted in figure 1b and figure 1c, where the latter gives the appearance of high rates for this area, probably due to the effect of environmental factors included in Model B, as compared to the municipalities aggregated in Model A.

In the maps with smoothed rates (figures 1b, 1c), the areas with higher admission rates are centered around municipalities with large industrial plants (such as Taranto in the Ionian curve) suggesting the effect of environmental risk factors and occupational exposure as determinants of higher rates of disease.

a) RAR before smoothing



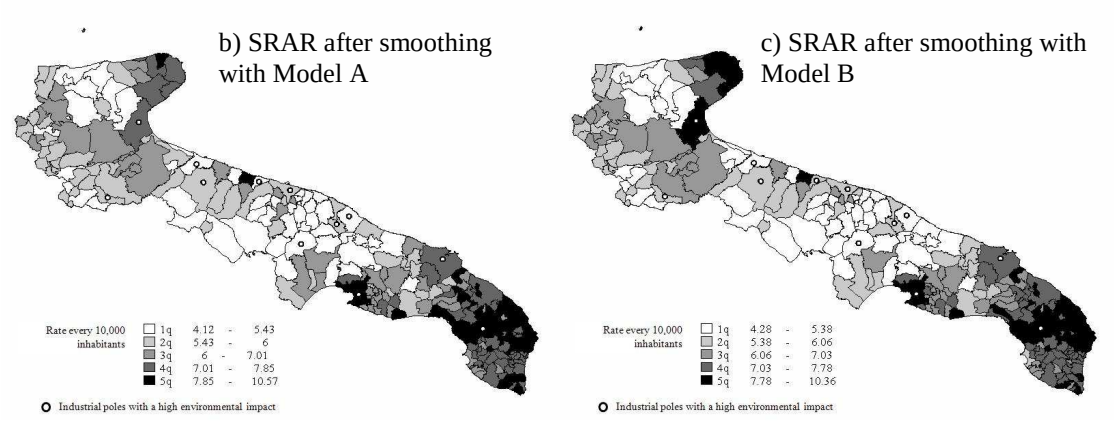


Figure 1: Maps of the Hospitalization Rate for Lung cancer. Apulia (Italy), 2006.

4. Concluding remarks

The results of the estimated models in which the clustering and heterogeneity components were adequately specified demonstrated that both heterogeneity and spatial autocorrelation were significant parameters. The effect of the smoothing procedure was greater in smaller municipalities, and especially in those with a more unstable RAR value. When the ASL was considered as a second hierarchical level parallel to that of spatial dependency, the municipalities heterogeneity component increased markedly and a better fit of the model to the data was obtained. The map of hospitalization rates for lung cancer in the Apulian Region estimated by the SRAR revealed the areas at higher risk better than the map estimated with the RAR. The inclusion of the ASL changed the spatial distribution of the risks, demonstrating a reduced hospitalization rate in the Gargano zone. This could probably be due to the different organization in this ASL, perhaps in the sense of a lesser likelihood of admitting patients to hospital and a lower availability or accessibility of diagnostic services, as compared with other ASL. The environmental risk, considered as a hierarchical level, did not provide a better explanation of the geographic distribution. Perhaps environmental risk should be entered in the model as a covariate, because it must be considered as an attribute of the municipality itself.

References

Olsen SF, Martuzzi M, Elliott P (1996) Cluster analysis and disease mapping-why, when and how?, *Br Med J*, 313, 863–866.

Carrington A, Heady P, Ralphs M, et al. (2007) Smoothing of Standardised Mortality Ratios. *Preliminary Investigation, National Statistics Methodological Series*, 35.

Fielding A, Goldstein H (2006) Cross-classified and Multiple Membership Structures in Multilevel Models: An Introduction and Review. *Research Report RR791. University of Birmingham*.

Goldstein H (2003) Multilevel Statistical Models, 3rd Edition. *London, Arnold*.

Goldstein H, Rasbash J, Plewis I, et al (1998) A User's Guide to MLwiN. *London: Institute of Education*.

Langford IH, Leyland AH, Rasbash J, Goldstein H (1999) Multilevel modelling of the geographical distribution of diseases. *J R Stat Soc Ser C Appl Stat*, 48(2), 253-268.

Dominici F, Peng RD, Bell ML, Pham L, McDermott A, Zeger SL, Samet JM (2006) Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA*, 295(10), 1127-1134.

A spatio-temporal model for air quality mapping using uncertain covariates [1]

Michela Cameletti
Università degli Studi di Bergamo

Stefania Ghigo, Rosaria Ignaccolo
Università degli Studi di Torino, ghigo@econ.unito.it

Abstract: Particulate matter (PM) is one of the most critical air pollutants because of its effects on the human health and the environment. It is well known that covariates, such as meteorological and geographical variables, have a significative influence on PM concentration. In this work we model PM concentration, measured by the monitoring network in Piemonte, taking into account the uncertainty of covariates that are output of a deterministic model chain, by means of a spatio-temporal error-in-variables model. The aim is to map the PM concentration random field all over Piemonte region considering all the uncertainty sources, i.e. the error related to the PM measurements and the covariate simulation as well as the error coming from the spatial prediction procedure.

Keywords: Error-in-variables model, Bayesian hierarchical model, MCMC

1 Introduction and motivating case study

The aim of this paper is to provide a spatio-temporal model of PM concentration, observed by a monitoring network, as function of some significative covariates (such as meteorological variables) given as output of a deterministic modeling system. While it is routine to consider that PM measurements are subject to an instrumental error, it is not usual to take into account the uncertainty of numerical model outputs. Usually such outputs are considered deterministic, thus known without error. However, numerical models try to reproduce reality but are affected by uncertainty related to initial conditions, parameters in model equations as well as model structure (Bayarri et al., 2009). To take into account these uncertainty sources we propose a spatio-temporal error-in-variables model (also known as measurement error model) where latent processes are introduced for modeling both the "true" PM and covariate fields. Our proposal is an extension of the models proposed in Van de Kassteele et al. (2006a, 2006b), where purely spatial error-in-variables models are considered in order to "correct" the numerical model outputs for nitrogen dioxide and particulate matter, respectively. Thus Van de Kassteele et al. (2006a, b) quantify the uncertainty of numerical model outputs, taking them as covariates in

[1] Work partially supported by Regione Piemonte.

a spatial model for the same pollutant. Instead, we want to take into account the uncertainty of exogenous covariates in air pollutant modelling.

In our case study, we consider daily particulate matter with an aerodynamic diameter of less than 10 μm (PM_{10}) measured at $n = 24$ sites and $T = 93$ days (from November 15, 2005 to February 15, 2006) in the Northern Italian region Piemonte. Moreover, we select $m = 10$ sites for validation purposes (see blue dots in Figure 1(a)). Because of the complex orography of the region, the pollutant dispersion is strongly affected by meteorological and geographical conditions. To take into account this relationship, we consider the following significative covariates (selected through a preliminary regression analysis): altitude (in m), coordinates (UTM, in km), daily mean wind speed (in m/s), daily mean temperature (in ^oK) daily maximum mixing height (in m) and daily emission rates of primary aerosols (in g/s). The time-varying covariates are simulated on a $4\,km \times 4\,km$ regular grid by a numerical model implemented by the environmental agency ARPA Piemonte (Bande et al., 2007) and are available at the monitoring sites as well. These numerical output covariates are introduced in our model with errors, whereas the constant in time covariates are supposed to be known without error.

2 The error-in-variables model

Let $y(s_i, t)$ and $x_k(s_i, t)$ denote, respectively, the measured PM_{10} concentration and the simulated value of the $k-$th covariate at location s_i and time t, with $i = 1, \ldots, n$, $t = 1, \ldots, T$ and $k = 1, \ldots, K$. Assuming that both $y(s_i, t)$ and $x_k(s_i, t)$ are affected by an additive error, we define the following equations

$$y(s_i, t) = \eta(s_i, t) + \varepsilon_y(s_i, t) \tag{1}$$

$$x_k(s_i, t) = \xi_k(s_i, t) + \varepsilon_{x_k}(s_i, t) \tag{2}$$

where $\eta(s_i, t)$ and $\xi_k(s_i, t)$ are two latent variables, $\varepsilon_y(s_i, t) \sim N(0, \sigma_y^2(s_i))$ and $\varepsilon_{x_k}(s_i, t) \sim N(0, \sigma_{x_k}^2(s_i))$ are the measurement and model errors, supposed to be independent. Moreover, we assume that the variances $\sigma_y^2(s_i)$ and $\sigma_{x_k}^2(s_i)$ do not depend on time and are known at each site s_i.

The relation between the two latent variables is defined by the following equation:

$$\eta(s_i, t) = \beta_0 + \boldsymbol{\gamma}_p \boldsymbol{z}(s_i) + \boldsymbol{\beta}_K \boldsymbol{\xi}(s_i, t) + \omega(s_i, t) + \varepsilon_q(s_i, t), \tag{3}$$

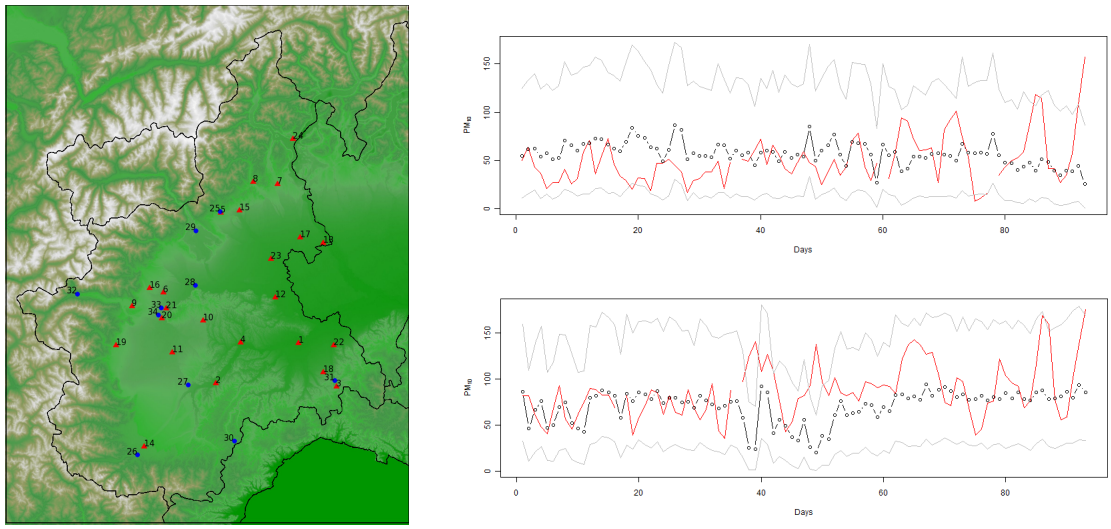
where $\boldsymbol{z}(s_i) = (z_1(s_i), \ldots, z_p(s_i))'$ is the vector of the p constant-in-time covariates known without error and $\boldsymbol{\gamma}_p = (\gamma_1, \ldots, \gamma_p)$ is the vector of their coefficients. Moreover, $\boldsymbol{\xi}(s_i, t) = (\xi_1(s_i, t), \ldots, \xi_K(s_i, t))'$ denotes the vector of the K "true" covariate values and $\boldsymbol{\beta}_K = (\beta_1, \ldots, \beta_K)$ is the vector of their coefficients. The term $\omega(s_i, t)$ is a spatio-temporal process assumed to be i.i.d. over time, so that the spatio-temporal covariance function is given by

$$Cov(\omega(s_i, t), \omega(s_j, t')) = \begin{cases} 0 & \text{if } t \neq t' \\ \sigma_\omega^2 \rho_\phi(h) & \text{if } t = t' \end{cases}$$

where $h = ||s_i - s_j||$ is the Euclidean distance between site s_i and s_j and σ^2_ω is the constant-in-time-and-space variance of the process. The function $\rho_\phi(h) = \exp(-\frac{h}{\phi})$ depends on the parameter ϕ, representing the decay rate of a spatial correlation with spatial distance. Finally, $\varepsilon_q(s_i, t)$ in Eq.(3) is the equation error that takes into account the not optimal relation between $\eta(s_i, t)$ and $\boldsymbol{\xi}(s_i, t)$; it is supposed to be normally distributed with zero mean and common variance σ^2_q. Thus, the parameter vector to be estimated is $\Phi = \{\beta_0, \boldsymbol{\gamma}_p, \boldsymbol{\beta}_K, \phi, \sigma^2_\omega, \sigma^2_q\}$. As regards inference, i.e. parameter estimation and spatial prediction of PM_{10} concentration at a new location s_0 and time t, we adopt a fully Bayesian framework via Markov chains Monte Carlo (MCMC) methods implemented through the `WinBUGS` software.

3 Results and concluding remarks

An exploratory analysis of the case study data showed skewed distributions for the considered variables. In order to make the PM_{10} and covariate distributions approximately Normal, a Box-Cox transformation (Box and Cox, 1964) was applied to the original data. Moreover, we standardized - site by site - the covariate data, in order to remove the effects related to the different ranges.



(a) PM_{10} site location

(b) Time series predictions

Figure 1: Locations of the 24 PM_{10} monitoring sites (red triangles) and 10 validation stations (blue dots) and prediction of PM_{10} for *26 Borgo San Dalmazzo* (top) and *28 Chivasso* (bottom) station: solid red line refers to PM_{10} observations, black dots to PM_{10} predictions and grey solid lines to 95% prediction intervals.

With regards to the variances supposed known in the model, in this preliminary study we fixed $\sigma^2_{x_k}(s_i) = 1$, $\forall i, k$ and $\sigma^2_y(s_i) = \sigma^2_y$ where σ^2_y is the variance of

PM_{10} data all over the sites. Considering the posterior estimates for the covariate coefficients, as expected there is a significative negative relationship between PM_{10} and altitude, as well as mean wind speed, mean temperature and maximum mixing height. The posterior mean of ϕ is 90.0423 which means that the spatial correlation decreases slowly with distance: for example, at 50 km the correlation is 0.5739 and 0.1212 at 190 km. Figure 1(b) displays the predicted PM_{10} for two different validation stations (*26 Borgo San Dalmazzo* and *28 Chivasso*). It seems that the predictions are close to the observed average for each of the ten sites, even though some problems can be detected when very high or very low PM_{10} concentration levels occur in contiguous days giving rise to a higher local variability. A possible solution to this issue can be achieved by choosing different values, one per site, of PM_{10} and covariate variances, in order to take into account the possibly different measurement error of PM_{10} and numerical model error of covariates in the sites.

Moreover, our ongoing research is focused on facing the so-called "change of support problem", which arises when the numerical model output is provided at a different spatial resolution from the scale of the PM measurements. Thus, it is interesting to extend the proposed spatio-temporal model in order to deal with both point-referenced and areal data.

References

Bande S., Clemente M., De Maria R., Muraro M., Picollo M., Arduino G., Calori G., Finardi S., Radice P., Silibello C., Brusasca G. (2007) The modelling system supporting Piemonte region yearly air quality assessment. *Proceedings of 6th International Conference on Urban Air Quality, Limassol*, Cyprus, 27-29 March 2007.

Bayarri M.J., Berger J., Steinberg D.M. (2009) Special Issue on Computer Modeling, *Technometrics*, 51(4), 353-353.

Box G.E.P., Cox D.R. (1964) An analysis of transformations, *Journal of the Royal Statistical Society, B*, 26, 211-246.

Van de Kassteele J., Stein A. (2006a) A model for external drift kriging with uncertain covariates applied to air quality measurements and dispersion model output, *Environmetrics*, 17, 309-322.

Van de Kassteele J., Koelemeijer R.B.A., Dekkers A.L.M., Schaap M., Homan C.D., Stein A. (2006b) Statistical mapping of PM_{10} concentrations over Western Europe using secondary information from dispersion modeling and MODIS satellite observations, *Stochastic Environmental Research and Risk Assessment*, 21, 2, 183-194.

African dust contribution on the PM10 daily exceedances occurred in Apulia region.

Lorenzo Angiuli [1], Roberto Giua [1], Simona Loguercio Polosa [1], Angela Morabito [1]
[1] Arpa Puglia, r.giua@arpa.puglia.it

Abstract: The air quality 2008/50/EC directive allows providing evidence of PM10 daily limit value exceedances due to natural sources, which are not to be considered for the purpose of the directive. In this work the African outbreaks, affecting the PM10 exceeding events occurred in Apulia region during 2010, are identified as follows. The PM10 daily concentrations were measured by the regional air quality monitoring stations, and complemented with meteorological maps, air mass back-trajectories, aerosol satellite retrievals, dust model simulations, ground measurements of aerosol optical properties. To quantify the daily net African dust load in PM10, we applied a methodology designed by the European Commission and based on the analysis of PM10 levels time series from regional background stations.

Keywords: mineral dust, Saharan outbreaks, PM10 daily limit value.

1. Introduction

The Italian peninsula, as well as the whole Mediterranean basin, is subjected to frequent Saharan dust events, especially during the summer season. The EU Air quality Directive 2008/50/CE allows providing evidences that the exceedances are due to natural sources. In this case, these exceedances are not considered as such for the purpose of the directive. In this study, for the 2010 main dust events occurred in Apulian region and identified on a daily resolution (Pederzoli et al. 2010), the daily net African dust load in PM10 has been quantified using a statistical methodology designed by the European Commission (Escudero et al., 2007). In this work we evaluated the occurrence of the PM10 exceedances, caused by African dust outbreaks in Apulia, and the mean annual contribution of African dust to PM10 levels at air quality monitoring sites.

2. Materials and Methods

The combined use of several sources of information allowed the identification of main Saharan dust outbreaks for Apulia region in 2010. A first discrimination of days with dust intrusions was carried out analyzing the daily series of PM10 concentrations as measured at *Cerrate* and *Ciuffreda*, the two regional background air quality monitoring stations. They are located in areas (figure 1a) far from urban and industrial areas so the anthropogenic contribution to PM10 concentration at these stations can be considered negligible. To confirm the saharan dust transport for each selected day, the 5-day isentropic back-trajectories at three different altitudes (750, 1500 and 2500 m.a.g.l.) were computed using HYSPLIT model with modeled vertical wind velocity. In addition, determination of the meteorological conditions causing the African dust outbreaks over Apulia region was carried out inspecting the NCEP synoptic

meteorological reanalysis during the period of dust outbreaks. The presence of dust over Apulia was also confirmed by looking at the maps of dust surface concentrations modeled by BSC-DREAM8b and by NAAPS global aerosol model. The satellite measurements have been used to detect the presence of atmospheric dust. In particular, the evolution of aerosol optical properties, like the Aerosol Optical Depth (AOD) and related Angstrom Exponent (α) at 550nm from Modis-Terra and Modis-Aqua satellite retrievals over the Mediterranean area, have been analyzed. Usually high AOD values combined with low α values are typical of Saharan dust (Pace et al., 2006). These values on Apulia region were also compared with the α and AOD values measured in the same days by an AERONET sun-photometer at Lecce.
The European Guidelines (Council of the European Union, 2011) propose a validated methodology for the quantification of the daily African PM10 load during dust outbreaks. This methodology is based on the subtraction of the daily regional background level from the PM10 concentration values at regional background stations. The daily regional background is obtained by computing a monthly 40th percentile to the PM10 time series at a regional background station, after a prior exclusion of the data of the days with African dust transport.

3. Results

In 2010 the PM10 exceedances due to Saharan dust outbreaks occurred on Apulia region only during winter and summer. The mechanism of mineral transport is different for the two seasons. In summer in North Africa the low precipitation level and the very high temperature are very favourable conditions for the massive resuspension of huge quantities and for the advection at different altitudes (up to 4-6km). In winter some sporadic Africans outbreaks may occur caused by depressions resulting in intense winds over the Saharan area. The description of the single summer event, identified by this methodology, is shown in the following figures. Figure 1b reports the daily series of PM10 concentrations measured at *Cerrate* and *Ciuffreda* between June 10th and June 22th. The data show an evident increase in the concentrations till June 16th. A similar trend was reported in all other PM10 monitoring stations.

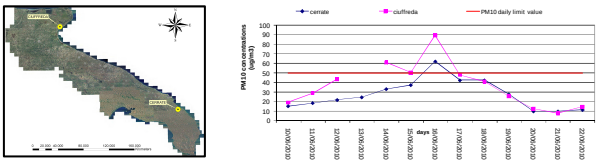


Figure 1 a) Rural background monitoring sites in Apulia region; b) Daily serie of PM10 concentration measured at Cerrate (LE) and at Ciuffreda (FG).

The meteorological analysis by the synoptic charts (not shown) has revealed for the 11th of June a 850 hPa african anticyclone on the North Africa, in gradual expansion on southern Italy until June 14th. In the following days this high-pressure system moves eastward continuing to affect Apulia up to June 20th, when the arrival of north Atlantic low pressure renews the air masses.
The transport of dust from Sahara up to June 20th is confirmed by a 5-day isentropic back-trajectories analysis computed at three different altitudes (750m, 1500m and 2500

m a.g.l.) using the HYSPLIT model. In figure 2a,b,c the back-trajectories with starting time 12 UTC on June14th, June 16th and June 20th are shown.

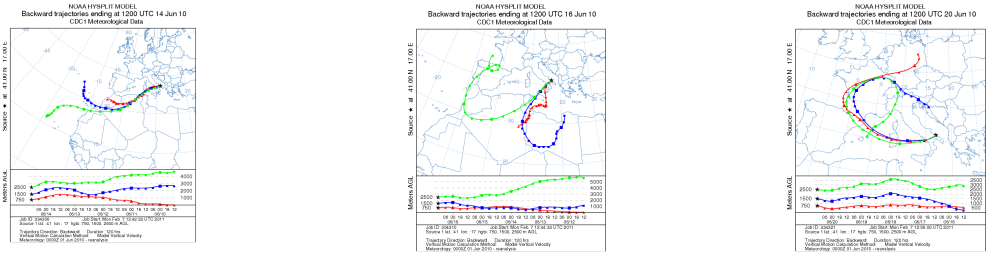


Figure 2 120-hours back-trajectories at 1200 UTC at 750m, 1500m, 2500m on a)14th,b)16th and c) 20th June .

The Saharan dust presence on Apulia is confirmed by satellite aerosol retrievals. Figure 3a,b,c shows the evolution of Aerosol Optical Depth (AOD) at 550 nm over the Mediterranean area, with grid resolution 1 degree x 1 degree, retrieved by MODIS-Terra for June 14th, 16th and 20th. Large amounts of dust (AOD>0.6) are observed as expected on June 14th and 16th, while on June 20th the AOD values are reduced significantly. Figure 4a,b shows the series of AOD and related α, as derived from observation by an AERONET sun-photometer at Lecce. AOD increases between June 11th and 20th, while the related Angstrom exponent (α) in the same period is reduced. Values of α greater than 2.0 indicate the presence of fine mode particles (e.g., smoke particles and sulfates), while values of α near zero indicate the presence of coarse mode particles such as desert dust. The Saharan dust event is also predicted by the BSC-DREAM8b model and by NAAPS Model. Maps of PM ground concentration in µg/m^3 from June 10th to 21th have been analyzed (not shown). The daily evolution of PM surface concentration over Apulia region, predicted by these models, is qualitatively quite similar to the trend of PM10 concentrations measured by air quality monitoring stations.

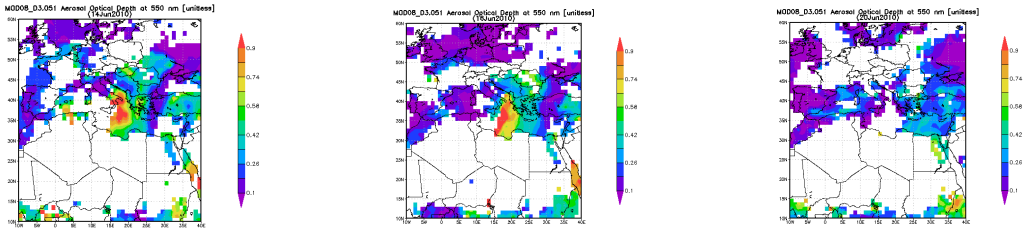


Figure 3 Maps of Aerosol Optical Depth (AOD) at 550 nm as retrieved by MODIS on a) June 14th, b) 16th and c) 20th.

Figure 4 Series of a) AOD at different wavelengths and Angstrom exponent measured

by sun photometer in June 2010. In Table 1 for each province of Apulia is reported the range of the occurrence of the PM10 daily exceedances and the range of the mean annual contribution to PM10 levels at air quality monitoring sites, caused by the main African dust outbreaks occurred during 2010.

Range of African dust contribution	BARI	BRINDISI	FOGGIA	LECCE	TARANTO
Occurrence of PM10 daily exceedances (%)	25 - 86	13 - 88	50 - 100	21 - 75	32 - 100
Mean annual of PM10 concentration ($\mu g/m^3$)	0.14 – 0.55	0.75 – 1.06	0.22- 0.43	0.67 – 0.98	0.60 – 1.13

Table 1 Range of the occurrence of the PM10 daily exceedances and range of the mean annual contribution to PM10 levels, for each Apulia province.

4. Concluding remarks

In this study we carried out i) the identification, on a daily resolution, of the 2010 main dust events occurred in Apulian region, ii) the quantification of PM10 exceedances caused by African dust outbreaks, iii) the mean annual contribution of African Dust to PM10 levels at air quality monitoring sites. In 2010 the PM10 exceedances due to Saharan dust outbreaks occurred on Apulia region only during winter and summer. As expected, the lowest percentages of African episodes are observed at monitoring stations where the local anthropogenic emissions (traffic, industrial, heating) are greater. In conclusion, in 2010 the mean annual contributions to PM10 levels were below or around $1 \mu g/m^3$.

References

Council of the European Union (February 2011), Commission staff working paper establishing guidelines for demonstration and subtraction of exceedances attributable to natural sources under the Directive 2008/50/EC on ambient air quality and cleaner air for Europe.

Escudero M., Querol X., Pey J., Alastuey A., Perez N.,Ferreira F.,Alonso S., Rodriguez S., Cuevas E. (2007) A methodology for the quantifications of the net Africam Dust load in air quality monitoring networks, *Atmospheric Environment*, 41, 5516-5524.

Pace G., di Sarra A., Meloni D., Piacentino S., Chamard P (2006) Aerosol optical properties at Lampedusa (Central Mediterranean). 1. Influence of transport and identification of different aerosol types, *Atmospheric Chemistry and Physic,s* 6, 697-713.

Pederzoli A., Mircea M., Finardi S., di Sarra A., Zanini G. (2010) Quantification of Saharan dust contribution to PM10 concentrations over Italy during 2003-2005, *Atmospheric Environment*, 44, 4181-4199.

Health impact assessment of pollution from incinerator in Modugno (Bari)

Ida Galise 1
Puglia Cancer Registry 1, ida.galise@oncologico.bari.it

Maria Serinelli 2
Regional Environment Protection Agency, Puglia, Bari, Italy 2

Lucia Bisceglia 1, 2
Puglia Cancer Registry 1
Regional Environment Protection Agency, Puglia, Bari, Italy 2

Giorgio Assennato 1, 2
Puglia Cancer Registry 1
Regional Environment Protection Agency, Puglia, Bari, Italy 2

Abstract: The purpose of this study is to assess the potential health impact at start-up of a new incinerator on the general population living near the facility. An algorithm was applied in order to calculate the number of deaths and hospital admissions associated with a given concentration of PM10, exposed population, specific mortality/morbidity rates. For every health end-points, an estimate of RR was obtained from literature. Using PM10 as a tracer, simulations were made of incinerator emissions fallout. Residents within 2-km radius from the plant were considered. The reduction of average concentration of PM10 to 40 $\mu g/m^3$ could prevent 0.12% of natural causes of death. Proportionally, the increment in PM10 concentration of 1 $\mu g/m^3$ could be associated to 0.02% of deaths. The estimated exposure to estimated incinerator emissions should not lead to additional health risks for the neighbouring population.

Keywords: air pollution, incinerator, hospital admissions, mortality.

1. Introduction

Several studies on the possible health effects related to population residing in the proximity of incinerators have been published, and well-conducted reviews are available on this subject. While some positive studies suggest associations with reproductive outcomes and cancer, the evidence is, overall, not conclusive to establish the occurrence and magnitude of risks. Furthermore, positive studies refer mostly to old generation incineration plants.
The adoption of Best Available Technologies (BAT) in abating emissions resulted in much lower levels of exposure to pollutants and consequently less likely occurrence of measurable health effects on populations resident in the proximity of new generation incinerators (Franchini et al. 2004, WHO 2007, Italian Epidemiological Association 2008, Porta et al. 2009, Ranzi et al. 2011).

This conclusion is supported mostly by extremely low concentrations of toxic substances measured in emissions of new generation incinerators (Moniter Projects Preliminary Results, 2010). However, residents' concerns -- living in areas near incinerators -- require evaluation of potential health effects associated to estimated emissions from new plants.
The aim of the present study is to assess the health impact on people living in the proximity of a new incineration facility in Modugno, Province of Bari (Puglia, Italy) in relation to PM10 exposure, by using current health records.

2. Materials and Methods

Health statistics on mortality for 2005 (last available year) were retrieved from the *Regional Mortality Atlas* (Regional Epidemiological Observatory, Puglia). The *Atlas* contains cause-specific mortality data at municipal level. Hospital Admissions (HA) data for this study were gathered from regional hospital discharge archives for 2008.
Mortality and morbidity end-points were chosen from the scientific evidence available and from recent evaluations of impact assessments. In particular, mortality endpoints include overall mortality (International Classification of Diseases, 9th Revision (ICD-9:1-799), cardiovascular (ICD 9: 390-459) and respiratory (ICD 9: 460-519) causes of death. Selected morbidity outcomes are related to cardiac (ICD 9: 390-429) and respiratory (ICD 9: 460-519) diseases. Hourly data on PM10 were obtained for the year 2008. The PM10 daily average, measured by air quality stations located in the area, was 45.3 μg/m^3. A simple algorithm was used to calculate the number of attributable deaths and hospital admissions associated with a given counterfactual factor of 40 μg/m^3 (as suggest from European Union limits), exposed population, specific mortality/morbidity rates and relative risk (RR) estimates (Martuzzi et al. 2006).
The number of cases attributable to an air pollution concentration over a given counterfactual factor, *E*, is given by the following equation:

$$E = A^* B^* (C/10)^* P,$$

where: *P* = the population exposed; *C* = the relevant change in concentration (difference between the observed concentration and the counterfactual level), obtained from monitoring networks in each city; *A* = the proportion of effect on health attributable to air pollution, which can be calculated as follows: $A = (RR-1)/RR$.
Residents living within 2 km from incinerator, area of expected maximal deposition estimated by ISAC-CNR-Lecce through dispersion modeling, were considered as exposed (15,056 inhabitants).
Concentration-response risk coefficients were derived from epidemiological studies (Table 1) (Martuzzi et al. 2006).

Outcomes	RR	CI 95%
All causes of mortality (excluding accidents)	1.006	1.004-1.008
Cardiovascular deaths	1.009	1.005-1.013
Respiratory deaths	1.013	1.005-1.020
Cardiac HA	1.003	1.000-1.006
Respiratory HA	1.006	1.002-1.011

Table 1: Summary of RRs and confidence interval 95% (95% CIs)

The analyses were performed in two steps:

- at first, we estimated how many deaths could have been avoided if the observed PM10 concentration could have been reduced to the given counterfactual level (40 $\mu g/m^3$);
- on a second phase, assuming that the incinerator will operate in combination with a Combined-Cycle combustion Gas Turbine (CCGT) power generation facility (that is another plant under construction in the area of study), we referred to the additional contribution to PM10 within 2 Km radius estimated through dispersion modeling: 0.03 $\mu g/m^3$ for incinerator and 0.15 $\mu g/m^3$ for CCGT plant. As worst-case scenario, we have chosen an increase of 1 $\mu g/m^3$ of PM10 exceeding 40 $\mu g/m^3$.

3. Results

The results of step 1 are reported in Tables 2 and 3. In detail, 0.12% of overall deaths, 0.19% of cardiovascular and 0.27% of respiratory mortality are attributable to levels of PM10 exceeding 40 $\mu g/m^3$ (Table 2). For morbidity: 0.06% of HA for cardiac and 0.12% of HA for respiratory diseases (Table 3).
The results of step 2 are reported in table 4.

Causes of death	Cases	Rates *(100,000 inhabitants)*	Attributable cases	CI 95%		% Attributable cases	CI 95%	
Overall mortality	208	562	0.26	0.17	0.34	0.12	0.08	0.17
Cardiovascular	78	211	0.15	0.08	0.21	0.19	0.11	0.28
Respiratory	18	49	0.05	0.02	0.08	0.27	0.11	0.42

Table 2: Cause-specific deaths attributable to mean levels of PM10 exceeding 40 $\mu g/m^3$. Modugno, 2008

Hospital admissions	Cases	Rates *(100,000 inhabitants)*	Attributable cases	CI 95%		% Attributable cases	CI 95%	
Cardiovascular	1,189	3.110	0.74	0.00	1.48	0.06	0.00	0.12
Respiratory	586	1.533	0.73	0.24	1.33	0.12	0.04	0.23

Table 3: Cause-specific hospital admissions attributable to mean levels of PM10 exceeding 40 $\mu g/m^3$. Modugno, 2008

Outcomes	Attributable cases	CI 95%		% Attributable cases	CI 95%	
Overall mortality	0.05	0.03	0.06	0.02	0.02	0.03
Cardiovascular mortality	0.03	0.02	0.04	0.04	0.02	0.05
Respiratory mortality	0.01	0.00	0.01	0.05	0.02	0.08
Cardiovascular HA	0.14	0.00	0.28	0.01	0.00	0.02
Respiratory HA	0.14	0.05	0.25	0.02	0.01	0.04

Table 4: Cases attributable to increase of 1 μg/m^3 of PM10 exceeding 40 μg/m^3. Modugno, 2008

4. Concluding remarks

Estimated PM10 levels associated to new incinerator emissions should not lead to additional health risks for the neighbouring population.
This evaluation is to be considered limited, given the following parameters: the small amount of residing population; the chemical and toxicological data of specific compounds; the characterization of individual exposure. Nevertheless, we confirm the need to activate an environmental and epidemiological surveillance in the examined area.

References

Franchini M., Rial M., Buiatti E., Bianchi F. (2004) Health effects of exposure to waste incinerator emissions: a review of epidemiological studies, *Ann. Ist. Super. Sanità*, 40:101-15.

WHO (2007) Population health and waste management: scientific data and policy options. Report of a WHO workshop, Rome, Italy. [http://www.euro.who.int/__data/assets/pdf_file/0012/91101/E91021.pdf].

Porta D., Milani S., Lazzarino A.I., Perucci C.A., Forastiere F. (2009) Systematic review of epidemiological studies on health effects associated with management of solid waste, *Environmental Health*, 8:60.

Ranzi A., Fano V., Erspamer L., Lauriola P., Perucci C.A., Forastiere F. (2011) Mortality and morbidity among people living close to incinerators: a cohort study based on dispersion modeling for exposure assessment, *Environmental Health*, 10:22.

AIE (2008) Waste processing and health. A position document of the Italian Association of Epidemiology (AIE), *Ann. Ist. Super. Sanità*, 44(3):301-306.

http://www.arpa.emr.it/cms3/documenti/moniter/risultati/20100914/02%20scheda_report_emissioni.pdf

Martuzzi M., Mitis F., Iavarone I., Serinelli M. (2006) Health impact assessment of PM10 and Ozone in 13 Italian cities. *World Health Organization - European Centre for Environment and Health.*

Local scoring rules for spatial processes

Philip Dawid
University of Cambridge <apd@statslab.cam.ac.uk>

Monica Musio
Università di Cagliari <mmusio@unica.it>

Abstract: We display pseudo-likelihood as a special case of a general estimation technique based on proper scoring rules. Such a rule supplies an unbiased estimating equation for any statistical model, and this can be extended to allow for missing data. When the scoring rule has a simple local structure, as in many spatial models, the need to compute problematic normalising constants is avoided. We illustrate the approach through an analysis of data on disease in bell pepper plants.

Keywords: proper scoring rule, pseudo-likelihood, ratio matching, unbiased estimating equation

1 Introduction

Maximum likelihood estimation of a spatial process can be computationally demanding because of the need to manipulate the normalisation constant of the joint distribution. Besag (1975) developed the method of *pseudo-likelihood* to sidestep this problem. This has traditionally been considered as an approximation (of unknown quality) to the full likelihood. However, as we describe below, the method can be justified in its own right, as leading to an unbiased estimating equation. Other methods, constructed from *proper scoring rules*, have similar justification and properties, and supply useful alternatives.

2 Proper scoring rules

A *scoring rule* $S(x, Q)$ is a loss function measuring the quality of a quoted probability distribution Q for a random variable X, in the light of the realised outcome x of X — see *e.g.* Dawid (1986). It is *proper* if, for any distribution P for X, the expected score $S(P, Q) := \mathrm{E}_{X\sim P} S(X, Q)$ is minimised by quoting $Q = P$. A prominent example is the *log score*, $-\log q(x)$, where q denotes the density or probability mass function of X.

Given a proper scoring rule S and a smooth parametric statistical model $\mathcal{P} = \{P_\theta\}$ for X, let

$$s(x, \theta) := \frac{\partial S(x, P_\theta)}{\partial \theta}.$$

Then we can estimate θ by $\widehat{\theta}_S$, the root of the *estimating equation*

$$s(x, \theta) = 0. \tag{1}$$

When S is the log score, this is just the likelihood equation, and $\widehat{\theta}_S$ is the maximum likelihood estimate. More generally, for any differentiable scoring rule and any smooth statistical model, $\mathrm{E}_\theta\{s(X, \theta)\} = 0$, *i.e.* (1) is an unbiased estimating equation (Dawid and Lauritzen 2005). In particular it will typically deliver a consistent, if not necessarily efficient, estimator in repeated sampling. We can then choose S to increase robustness or ease of computation.

In the context of a spatial process $X = (X_v : v \in V)$, we can define a useful class of proper scoring rules (Dawid *et al.* 2011) by

$$S(x, Q) = \sum_v S_0(x_v, Q_v), \tag{2}$$

where Q_v is the conditional distribution of X_v, given the values $x_{\setminus v}$ for the variables $X_{\setminus v}$ at all sites other than V, and S_0 is a proper scoring rule for the state at a single site. In particular, if Q is Markov on a graph $\mathcal{G}$, then Q_v only depends on the values $x_{\mathrm{ne}(v)}$ at the sites neighbouring v. This avoids the need to evaluate the normalising constant of the full joint distribution Q.

Corresponding to (2) we have estimating equation

$$\sum_v s_0(x_v, P_{\theta,v}) = 0 \tag{3}$$

with each term in the sum having expectation 0. When S_0 is the log score, (3) gives the (negative log) *pseudo-likelihood* (Besag 1975). For X_v binary and S_0 the quadratic ("Brier") score, it yields the method of *ratio matching* (Hyvärinen 2007).

Missing data are readily dealt with (although with some loss of efficiency). Let $A_v = 1$ if any value in $\{v\} \cup \mathrm{ne}(v)$ is missing. Then so long as the data are missing completely at random, $s_0(x_v, P_{\theta,v}) \times A_v$ has expectation 0, so we can just omit incomplete terms from (3) while retaining an unbiased estimating equation.

3 Phytophthora data

Figure 1 displays the presence or absence of the pathogen *Phytophthora capsici* Leonian in bell pepper plants on a regular 20×20 grid (Chadoeuf *et al.* 1992).

We model the data as a stationary first-order Markov process with respect to the grid, which thus follows the autologistic model (Besag 1972; Besag 1974; Gumpertz *et al.* 1997):

$$\operatorname{logit} \pi_{ij} = \alpha + \beta(x_{i-1,j} + x_{i+1,j}) + \gamma(x_{i,j-1} + x_{i,j+1}) \tag{4}$$

where π_{ij} is the probability of $X_{ij} = 1$, given all other values.

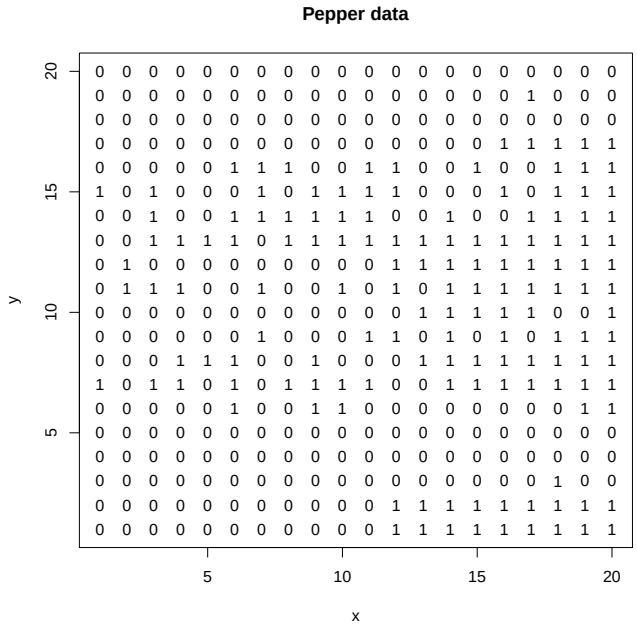


Figure 1: Presence (1) and absence (0) of pathogen in bell pepper plants

To fit by maximum pseudo-likelihood (PL), we simply proceed as if the (X_{ij}) were all independent, and maximise the resulting "likelihood". This can be done by a standard generalized linear model analysis, readily implemented in standard software such as `R`, using the binomial family and (default) logit link function.

Alternatively, and possibly more robustly, we could apply ratio matching (RM), based on the Brier scoring rule, which leads to the least-squares recipe: minimise $\sum(x_{ij} - \pi_{ij})^2$. Again this can be implemented in standard GLM software, treating the data as if they were normal with constant variance, and using the logit link function (in `R` this is effected using the `glm()` command with option `family=quasi(link=logit,variance=constant).`)

Note however that, although it is easy to compute the estimates, the associated "standard errors" output by the software will be inappropriate, since they do not take account of the dependence in the data.

4 Results

Table 1 displays the results of fitting the model (4) by pseudo-likelihood (PS) and by ratio matching (RM). Values at sites on the boundary of the grid, which do not have four observed neighbours, are not used as responses, though they are used as covariate values for their neighbouring interior sites. There are thus $18 \times 18 = 324$ data-points used to fit the model.

Method	Intercept, α	WE, β	NS, γ
PL	-2.4390	1.6514	0.6266
RM	-2.2654	1.5864	0.5375

Table 1: Coefficients estimated by pseudo-likelihood (PS) and ratio matching (RM)

5 Concluding remarks

The PL and RM methods, as well as others derived from different proper scoring rules, all involve solving an unbiased estimating equation. In the example studied, the estimates from PL and RM are broadly in line. However further theoretical and experimental work is needed to explore and compare their accuracy, efficiency and robustness properties.

References

Besag, J. E. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 75–83.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, **36**, (2), 192–236.

Besag, J. E. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **24**, 179–95.

Chadoeuf, J., Nandris, D., Geiger, J., Nicole, M., and Pierrat, J. (1992). Modelisation spatio-temporelle d'une epidémie par un processus de Gibbs: Estimation et tests. *Biometrics*, **48**, 1165–75.

Dawid, A. P. (1986). Probability forecasting. In *Encyclopedia of Statistical Sciences*, (ed. S. Kotz, N. L. Johnson, and C. B. Read), pp. 210–8. Wiley-Interscience.

Dawid, A. P. and Lauritzen, S. L. (2005). The geometry of decision theory. In *Proceedings of the Second International Symposium on Information Geometry and its Applications*, pp. 22–8. University of Tokyo.

Dawid, A. P., Lauritzen, S. L., and Parry, M. (2011). Proper scoring rules on discrete sample spaces. `arXiv:1104.2224`.

Gumpertz, M. L., Graham, J. M., and Ristaino, J. B. (1997). Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: Effects of soil variables on disease presence. *Journal of Agricultural, Biological, and Environmental Statistics*, **2**, 131–56.

Hyvärinen, A. (2007). Some extensions of score matching. *Computational Statistics and Data Analysis*, **51**, 2499–512.

Measuring Urban Quality of Life Using Multivariate Geostatistical Models [1]

Alessandra Michelangeli
University of Milan-Bicocca, alessandra.michelangeli@unimib.it
Clarissa Ferrari, Marco Minozzo
University of Verona

Abstract: Urban quality of life (QOL) is usually measured through an index defined as the estimated value of a set of urban amenities. However there is an increasing awareness that omitted variables might seriously undermine the method's ability to accurately estimate QOL. Here we extend the hedonic approach using a multivariate geostatistical model to address the omitted variable bias by identifying the latent common factors responsible for the spatial distribution of the amenities. A new QOL index is then defined as a function of the latent factors whose implicit prices are estimated through hedonic regressions. Our methodology is shown on a data set of individual-level property transactions from the city of Vicenza. As a result we obtain the spatial distribution of QOL calculated according to the new index.

Keywords: Hedonic prices, Housing market, Monte Carlo EM algorithm, Spatial factor model.

1 Introduction

The well-being of people living in a city depends on the level of development of the city itself. The more the city is able to provide services and infrastructures, the better the living conditions are for its inhabitants. Obviously the range of factors that affect QOL is much wider and includes, among the others, climate conditions, environmental quality, the level of security for persons and things, and the socio-demographic environment. However, only partial statistical information is generally available, concerning a limited set of indicators that do not exhaust all the relevant factors. In this paper, we formulate a specific hypothesis to overcome the problem of shortage of statistical information. We suppose that the diverse factors affecting QOL can be subdivided into two groups; factors depending on the intervention of the public authority or of a private agent constitute the 'Agent-dependent factors group', whereas factors that do not depend on the intervention of some agent, as some characteristics of the landscape, constitute the 'Agent-independent factors group'. Then, assuming a model similar to that of Minozzo and Fruttini (2004), the

[1] We gratefully acknowledge funding from the Italian Ministry of Education, University and Research (MIUR) through PRIN 2008 project 2008MRFM2H, and Polo Scientifico Didattico "Studi sull'Impresa" (Vicenza) of the University of Verona.

two latent common factors behind these two groups are identified. Once the spatial distributions of these two latent factors are evaluated, we are able to assess their impact on QOL following the hedonic theoretical model of Rosen (1979)and Roback (1982). This approach usually defines a QOL index as a weighted sum of a set of urban amenities, where the weights are the hedonic prices of the amenities derived from the compensating differentials in the housing or in the labor markets, or in both.

We proceed as follows: in Section 2, we describe the different sources of available data. In Section 3 we briefly review the theoretical framework developed by Rosen (1979) and Roback (1982) from which the QOL is recovered. Then we define the multivariate geostatistical model used to identify the latent factors. Finally, we define a new QOL index as a function of the latent factors. The last section concludes the work.

2 Housing and Urban Data

The data come from different sources and are combined into a single data set. Housing market data come from the 'Osservatorio del Mercato Immobiliare' (OMI) managed by a public agency ('Agenzia del Territorio'), and refer to some 600 individual house transactions in Vicenza between 2004 and 2009. In addition to housing market values, the data set provides information also on structural characteristics of the properties. Housing prices are expressed in 2004 constant euros. On the other hand, the data on the amenities and the socio-demographic characteristics of the city are from the municipality and refer to environmental characteristics, educational services, commercial and administrative facilities, and public transports.All housing units and local amenities were geocoded by assigning to each of them a latitude and longitude coordinate by using a GIS-based geocoding application. Then, for each of the K housing units, that is, for each pointwise geographic location $\mathbf{x}_k$, $k = 1, \ldots, K$, of these units, we computed the Euclidean distance from the unit to the nearest representative of each of m categories of amenities. These distances, which we indicate with $y_1(\mathbf{x}_k), \ldots, y_m(\mathbf{x}_k)$, for $k = 1, \ldots, K$, constitute the key data for our geostatistical factor model.

3 Detection of Spatial Latent Factors for QOL

Our methodology is based on the model developed by Rosen (1979) and by Roback (1982) to assess urban QOL. The model depicts cities as interrelated bundles of wages, rents and amenities, with the specific combination of these elements differing across cities. Households and firms choose their location to, respectively, maximize utility and minimize production costs. Households are assumed as workers which are homogeneous in income and tastes. They maximize their utility function choosing the optimal bundle of composite good and residential location which allows access to

local amenities. Firms combine capital and local labor using a production technology with constant returns to scale. The cost function depends both on input prices and the bundles of amenities. In equilibrium all households, regardless of their location, attain a common level of utility, and the unit production costs are equal to the unit production price. Equilibrium differentials for wages and housing prices can be used to compute implicit prices of amenities. Given the estimates of the implicit prices of amenities, the QOL index for any urban area is obtained by summing over all the average quantities of amenities using the implicit prices as weights. A serious drawback of this model is that the information about all urban attributes affecting QOL is unlikely to be available, and as Blomquist et al. (1988) note, even if all data were available, econometric specification problems such as collinearity would prevent the inclusion of all urban amenities. Furthermore, economic theory does not provide guidance for determining the optimal list of attributes. As a result, the empirical specification of the model may be plagued by omitted variable bias and measurement errors.

We overcome these problems by complementing this hedonic approach with the use of a hierarchical geostatistical factor model which will allow to arrive at a new QOL index. This model, which allows to deal with non-Gaussian data, can be seen as an extension to the multivariate context, of the classical geostatistical linear model of coregionalization (or of the spatial generalized linear model (Wang and Wall, 2003)). For a given set of distances $y_1(\mathbf{x}_k), \ldots, y_m(\mathbf{x}_k)$, for $k = 1, \ldots, K$, we assume that these are the realization, at the set of spatial locations $\mathbf{x}_1, \ldots, \mathbf{x}_K$, of a set of m random functions $Y_1, \ldots, Y_m$, for which we assume that, for $\mathbf{x} \in \mathbf{R}^2$,

$$Y_i(\mathbf{x})|Z_i(\mathbf{x}) \sim f_i(y; M_i(\mathbf{x})),$$

and

$$Z_i(\mathbf{x}) = \sum_{p=1}^{P} a_{ip} F_p(\mathbf{x}) + \xi_i(\mathbf{x}),$$

where $f_i(y; M_i(\mathbf{x}))$ is a Gamma density and $M_i(\mathbf{x}) = \mathrm{E}[Y_i(\mathbf{x})|Z_i(\mathbf{x})]$ is the conditional mean of the data given the latent part of the model, which is linked to $Z_i(\mathbf{x})$ by the link function $h_i(M_i(\mathbf{x})) = \beta_i + Z_i(\mathbf{x})$. As we said, $Y_i(\mathbf{x}_k)$ represents the minimum distance between the housing unit located at point $\mathbf{x}_k$ and the ith amenity. The latent part of the model resembles the classical linear factor model. Here, $F_p(\mathbf{x})$, for $p = 1, \ldots, P$, and $\xi_i(\mathbf{x})$, for $i = 1, \ldots, m$ are the common and the unique factors of the model, that we assume Gaussian, and for which we assume a spatial autocorrelation structure depending on a spatial autocorrelation function $\rho(\mathbf{h})$, for $\mathbf{h} \in \mathbf{R}^2$, such that $\rho(0) = 1$, and $\rho(\mathbf{h}) \to 0$, as $|\mathbf{h}| \to \infty$. In particular, we assume that $\mathrm{Cov}[F_p(\mathbf{x}), F_p(\mathbf{x} + \mathbf{h})] = \rho(\mathbf{h})$, and $\mathrm{Cov}[\xi_i(\mathbf{x}), \xi_i(\mathbf{x} + \mathbf{h})] = \psi_i \rho(\mathbf{h})$, for $\psi_i > 0$ (for more details see Minozzo and Fruttini, 2004, Minozzo and Ferrari, 2010). According to our proposal, we assume here that $P = 2$, that is, that there are two common latent factors responsible for the spatial distribution of the two groups of urban amenities ('Agent-dependent' and 'Agent-independent'). For this model,

the intercept parameters $\beta_1, \ldots, \beta_m$, the coefficients a_{i1}, a_{i2}, for $i = 1, \ldots, m$, and the variances $\psi_1, \ldots, \psi_m$, can be estimated through Monte Carlo EM algorithms, whereas predictions of the common latent factors $F_1(\mathbf{x}_k)$ and $F_2(\mathbf{x}_k)$, at the spatial locations $\mathbf{x}_k$, for $k = 1, \ldots, K$, can be obtained by MCMC methods. Once the predicted spatial distributions of the two common latent factors have been obtained over the city area, and in particular at the set of K property locations, we can proceed in using this predictions to obtain the spatial distribution of our QOL index, defined as the weighted sum of the two common latent factors, using the implicit prices as weights. These implicit prices can be estimated by simple hedonic regressions or adopting more sophisticated techniques.

Preliminary assessments confirm a monocentric structure of the city QOL is higher in the city centre where there is a greater quantity and variety of amenities and decreases non monotonically as the distance from the center increases.

4 Concluding remarks

This paper extends the hedonic approach to measure urban QOL by using a hierarchical geostatistical factor model. The proposed methodology improves our ability to assess the spatial distribution of QOL and to identify its main causes. The standard approach developed by Rosen (1979) and Roback (1982) gives a synthetic measure of the QOL that people can on average enjoy in the considered areal unit, for example a city or a neighbourhood. Our approach allows to determine the QOL level at each spatial point obtaining the entire spatial distribution of QOL index instead of its average value.

References

Roback J. (1982) Wages, rents, and the quality of life, *Journal of Political Economy* University of Chicago Press, 90, 1257–1278.

Rosen S. (1979) Wage-based indexes of urban quality of life, in *Current Issues in Urban Economics*, Mieszkowsi, P. & Stratzheim, M. (Eds.), John Hopkins Press, Baltimore, 74–104.

Minozzo M., Ferrari C. (2010) A hierarchical geostatistical factor model for multivariate Poisson count data, *Working Paper Series, Department of Economics, University of Verona.*

Minozzo M., Fruttini D. (2004) Loglinear spatial factor analysis: an application to diabetes mellitus complications, *Environmetrics*, 15, 423–434.

Wang F., Wall M. M. (2003) Generalized common spatial factor model, *Biostatistics*, 4, 569–582.

Multivariate and Spatial Extremes for the Analysis of Air Quality Data [1]

Simone A. Padoan and Alessandro Fassó
Department of Information Technology and Mathematical Methods,
University of Bergamo, Viale Marconi 5, 24044 Dalmine, Bergamo, Italy
email: simone.padoan@unibg.

Abstract: In recent years statistical analyses for monitoring the environment are increasingly in demand in different areas such as epidemiology, engineering, economy, etc. An example is the statistical monitoring of air quality, which makes it possible to statistically quantify the amount of certain pollutants in the lower troposphere. For a better understanding of the stochastic behavior of pollutants we focus on describing their extreme responses, because excessively extreme levels in the air may have implications in the environment and on human health. We then consider multivariate extreme value models and the class of maxstable processes in order to asses the frequencies of several extreme pollutant levels in central Europe and their spatial dependence structure.

Keywords: max-stable processes, multivariate extreme value distributions, generalized extreme value distribution, extremal coefficient, correlation function, Fréchet distribution, pollution.

1 Introduction

Nowadays in many disciplines such as epidemiology, engineering, economy, etc, are in great demand the statistical analyses for monitoring the environment. Specifically, it is very important to statistically quantify the amount of certain pollutants in the lower troposphere and this is possible thanks to the statistical monitoring of the air quality. A main aspect of environmental processes is their natural spatial domain, presupposing a statistical spatial analysis approach. One of the primary aims of the latter is to asses the dependence structure of the underlying process. In this case it is important to determine the degree of dependence of the pollutants' levels among the monitoring stations. There are a number of generic approaches to spatial modeling that to date have already been widely applied (e.g., Diggle and Ribeiro, 2007). But these are suitable for modeling the mean process levels, therefore they are inappropriate for handling extremal aspects. For a better understanding of the stochastic behavior of pollutants we focus on describing their extreme responses, be-

[1]This research is part of Project EN17, "Methods for the integration of different renewable energy sources and impact monitoring with satellite data", funded by Lombardy Region under "Frame Agreement 2009".

cause excessively extreme levels in the air may have implications in the environment and on human health.

With this work we aim to describe the extreme values of certain pollutants, such as fine particulate matters, sulphure, nitrogen dioxides, etc. recorded in central Europe. Each pollutant is recorded at $s = 1, \ldots, S$ locations, within a continuous region, for n-temporal observation with $n = 1, 2, \ldots$. At each site we compute the maximum with respect to a block of N temporal observations. For example, for hourly observations, we set $N = 24 \times 366$ and this implies that we focus on annual maxima of the process. Thus, we derive a temporal series of componentwise maxima of process measurements denoted by $\{y_t(s)\}$ with $t = 1, \ldots, T$ the sample of block maxima. In order to perform the analyses of the pollutants' extreme levels we consider the classes of multivariate extreme value models and of maxstable processes (see e.g. Chapters 6, 9 of de Haan and Ferreira, 2006). These families provide a quite general framework, with similar asymptotic motivations to the univariate case, suitable to model extreme processes incorporating temporal or spatial dependence. Statistical methods for max-stable processes and data analyses of practical problems are discussed by Padoan et al. (2010).

2 Methods

A suitable setting for addressing spatial problems in the extreme values context is provided by max-stable processes.

Let $\{Y(x)\}_{x\in\mathcal{X}}$ be a stochastic process defined on $\mathcal{X} \subseteq \mathbb{R}^q$, $q \in \mathbb{N}$, with continuous sample path. Assume that n independent and identically distributed (iid) copies of it, Y_i with $i = 1, \ldots, n$, are available, and hence focus on the limit of the rescaled process $\{M_n(x)\}_{x\in\mathcal{X}}$. Specifically, if there exist continuous positive functions $a_n(x)$ and real functions $b_n(x)$, with $n \in \mathbb{N}$ such that

$$Z(x) = \lim_{n\to\infty} \left\{ \frac{M_n(x) - b_n(x)}{a_n(x)} \right\}_{x\in\mathcal{X}} \tag{1}$$

is not a trivial limit, that is the normalized sequence $M_n(x)$ converges in distribution to a process $Z(x)$ with non-degenerate marginals for all $x \in \mathcal{X}$, then we call Z an extreme value process. Observe, that the limiting process Z posses three important proprieties: a) it is a *max-stable* process; b) all its univariate marginal distributions belong to *the generalized extreme value* class of distributions; c) all its finite p-dimensional distributions, with $p \geq 2$, are characterized to be *multivariate extreme value distributions* (see e.g. Chapters 1, 6 of de Haan and Ferreria, 2006).

Correlation coefficients and correlation functions are typically used in order to describe pairwise dependence, under Gaussianity assumption, respectively for high dimensional and spatial analysis. Similarly extremal coefficients and the extremal coefficient functions describe the dependence for extremes. Specifically, given Z_i, $i = 1, \ldots, n$, iid copies of a component-wise random vector $Z = (Z_1, \ldots, Z_p) \in \mathbb{R}_+^p$

with common unit Fréchet margins, then from the following relation

$$\mathbb{P}\{\max(Z_1,\ldots,Z_p)\leq z\}=\mathbb{P}\{Z_1\leq z\}^{\theta}=\exp(-\theta/z),\quad z>0,$$

where the rightmost term is a Fréchet(θ) distribution, the parameter $1\leq\theta\leq p$ defines the extremal coefficient. When $\theta=1$ indicates complete dependence, whereas $\theta=p$ corresponds to full independence. The extremal dependence of stochastic processes has a similar definition. If now we consider a stationary max-stable process $Z(x)$ with univariate unit Fréchet margins then, for any pair of locations x_1, $x_2\in\mathcal{X}$ separated by $h=x_2-x_1$, from the following relation

$$\mathbb{P}\{\max(Z(h),Z(o))\leq z\}=\exp(-\theta(h)/z),\quad z>0,$$

the real-valued function $\theta(h)$ defines the pairwise extremal coefficient function, where o denotes the origin (e.g. Schlather and Tawn, 2003). From a practical point of view we consider, for modeling the extremes of the pollutants, two specific families of max-stable processes such as the Brown-Resnick process (e.g. Kabluchko et al., 2009) and the Extremal Gausssian process (e.g. Schlather, 2002) and the class of multivariate extreme value distributions named the Extremal-t model (e.g. Nikoloulopoulos et al., 2009). We can easily fit these models to the pollutants data using the maximum composite likelihood estimation method (e.g. Padoan et al., 2010) and then to compare the different results. Moreover, for these models the closed form of the extremal coefficients is known so that we can, after the fitting step, assess the dependence structure and estimate the frequencies with which different high levels of the pollutants occur.

3 Data

The dataset considered for the analysis consists of hourly measurements of some pollutants of central Europe (see left panel of Figure 1) available on the Internet at the website: `http://www.eea.europa.eu`. Specifically, we took into account a time-period of 13 years, from January 1996 to December 2008, and we selected a region of approximately 341.000 km^2. The right panel of Figure 1 shows the area where the monitoring weather stations are located and the numbers from 1 to 3 denote the locations for the different pollutants. In particular, the 139 stations indicated with the number 1 monitor the benzene ($COH6$), carbon monoxide (CO) and nitrogen dioxide ($NO2$), the 126 stations indicated by the number 2 monitor the ozone ($O3$) and the 68 stations indicated by the number 3 monitor the particulate matters ($PM10$), sulphure ($SO2$). For each pollutant there are some missing data but the percentage is small, we can account on average (between sites) 2 % of missing values. Given that in the analysis we focus on block maxima of pollutants, where the blocks are formed by 24×366 temporal observations leading to sequences of annual maxima, the small percentage of missing data should not have an impact on the description of the extremes.

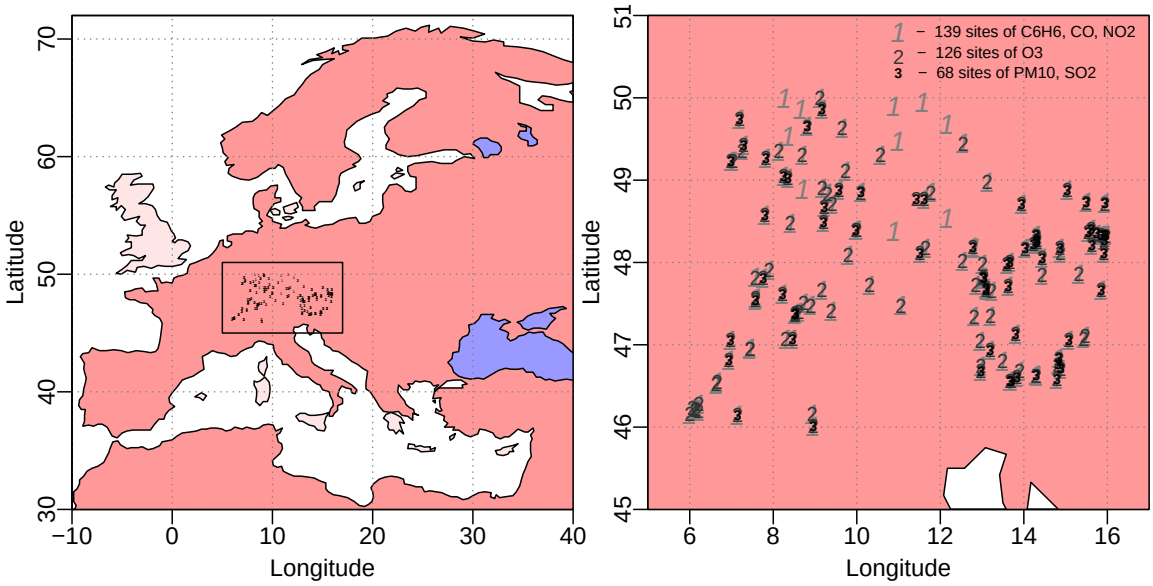


Figure 1: *Air quality data: the left panel reports the European map and the rectangle displays the central part where the monitoring weather stations are located. The right panel shows the expanded zone marked by the rectangle of the left panel and displays with the numbers the locations of the monitoring stations.*

References

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory An Introduction*. New York: Springer.

Kabluchko, Z., Schlather, M. and de Haan, L. (2009). Stationary max-stable fields associated to negative definite functions. *The Annals of Probability*, 37, 2042–2065.

Nikoloulopoulos, A. K., Joe H. and Li H. (2009). Extreme value properties of multivariate t copulas. *Extremes*, 12, 129–148.

Padoan, S. A., Ribatet, M. and Sisson S. A. (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association, Theory & Methods*, 105, 263–277.

Diggle, P. J. and Ribeiro P. J. (2007) *Model–Based Geostatistics*. London: Springer.

Schlather, M. (2002). Models for stationary max-stable random fields. *Extremes*, 5, 33–44.

Schlather, M. and J. A. Tawn (2003). A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika*, 90, 139–154.

Pulmonary Tuberculosis and HIV/AIDS in Portugal: joint spatiotemporal clustering under an epidemiological perspective [1]

Carla Nunes and Teodoro Briz
CIESP/ENSP and CMDT.LA, UNL cnunes@ensp.unl.pt

Dulce Gomes and Patrícia A. Filipe
CIMA, Universidade de Évora

Abstract: Since the mid-80s Tuberculosis declining trend became softer and even reversed in some countries. HIV/AIDS frequently appears as the main cause for the resurgence of Tuberculosis. This work aims at identifying critical areas for the joint occurrence of these conditions in Portugal, and at confirming the belief that HIV is not a major explanation for the slow Tuberculosis incidence decline.
Based on correlation analyses and space-time scan statistics, a weak statistical correlation between HIV and TB incidence rates were observed (0.279; $p<0.001$). For both diseases, Oporto and Lisbon Metropolitan Areas were identified as critical locals, with relative risks of, respectively, 1.77 and 1.78 for TB, and 5.66 and 3.31 for HIV. Similar areas were identified with a multivariate scan.
Keywords: Tuberculosis, HIV/AIDS, Spatiotemporal clustering,

1. Introduction

Although Tuberculosis and HIV /AIDS are presenting a decreasing trend, especially in developed countries, critical areas must be identified, to allow a global control of both diseases (Anandaiah *et al.*, 2011; de Colombani *et al*., 2004). Also, it is well known and stated in scientific literature that the particular and strong relation between these two diseases encourages joint actions (Bhagyabati Devi *et al.*, 2005; Couceiro *et al*., 2011). Some epidemiological dimensions (e.g.: risk factors) that are common to both TB and HIV lead to challenges and opportunities in surveillance, by promoting joint surveillance and control programs, including the development of TB/HIV indicators (Sanchez *et al.*, 2010). Only pulmonary cases are considered, given their outstanding role in disease transmission (Couceiro *et al*., 2011).
The goals of this study are: to identify critical areas for each disease separately, based on two independent datasets, in Portugal Mainland, per municipality and per year (2000-2009); and to identify and characterize critical areas for the joint occurrence of both diseases.

2. Materials and Methods

Data available concern the period 2000 to 2009, per municipality and were provided by three different official sources: the number of Tuberculosis notified cases - by the National Program for Tuberculosis Control; the number of HIV/AIDS notified cases - by the National Registry of HIV infected individuals (Communicable Diseases and

[1] CIMA/EU, and Lilly Portugal - Produtos Farmacêuticos, Lda. from Project: A Tuberculose em Portugal e seus determinantes.

Epidemiological Surveillance Center, INSA); and population data - by the National Statistical Institute. Constant detection rates among municipalities and in time are assumed. This study is exploratory and the interference of external relevant dimensions in the focused associations was not accounted for yet.

In a first approach, for either HIV/AIDS or Tuberculosis incidence rates, some independent exploratory data analyses were conducted. Additionally, correlation analyses pertaining to both entities were done.

In order to identify critical high incidence areas, spatiotemporal clustering analyses, based on the space-time scan statistic (Kulldorff, 1997), were applied separately for each disease. A multivariate scan with a multiple data sets process was applied, as HIV/AIDS is one of the risk factors for Tuberculosis and because both diseases have common risk factors. This method allows us to identify significant space-time joint clusters.

Space-time scan statistic is one of the most referred techniques in spatiotemporal epidemiological scientific literature, due to its appropriateness for the purpose, to its robust theoretical framework and also to the existence of free and friendly software (www.satscan.org).

3. Results

Brief descriptive analyses of Pulmonary Tuberculosis incidence rates (TBIR) and of HIV/AIDS incidence rates (HIVIR), based on incidence rates per municipality and per year (2000-2009), are presented in Tables 1 and 2 and Figure 1.

Table 1. Municipalities TBIR Descriptive Statistics (10^{-5}), global and per year.

TBIR	Global	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Mean	20.50	24.59	22.57	23.25	21.43	20.01	20.11	18.13	16.09	20.16	18.71
Median	16.99	20.73	18.15	19.34	18.02	16.63	17.50	14.77	13.95	16.53	16.24
Std. Deviation	18.51	22.58	19.80	19.44	18.13	17.98	17.57	16.45	14.76	17.92	18.08
Maximum	155.70	144.58	109.57	92.87	103.14	112.28	90.80	77.39	68.98	95.02	155.70

A slow, hesitating and declining global trend in time was apparent for mean TBIR in the Mainland, between 2000 and 2009 (Table 1). The global mean (standard deviation) for TBIR was 20.50 (18.51).

Table 2. Municipalities HIVIR Descriptive Statistics (10^{-5}), global and per year.

HIVIR	Global	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Mean	7.87	9.73	9.64	8.88	9.18	8.04	7.60	7.89	7.37	6.37	4.02
Median	2.57	2.39	3.98	4.43	5.42	2.36	4.03	4.22	3.06	.00	.00
Std. Deviation	12.73	18.04	14.69	13.59	12.79	12.30	11.33	11.18	11.73	10.69	7.26
Maximum	192.07	192.07	106.54	104.83	103.27	105.04	96.37	82.53	87.92	77.75	43.92

A steeper global decline of HIVIR was observed, in the same period and area (Table 2), with a global mean (standard deviation) of 7.87 (12.73).

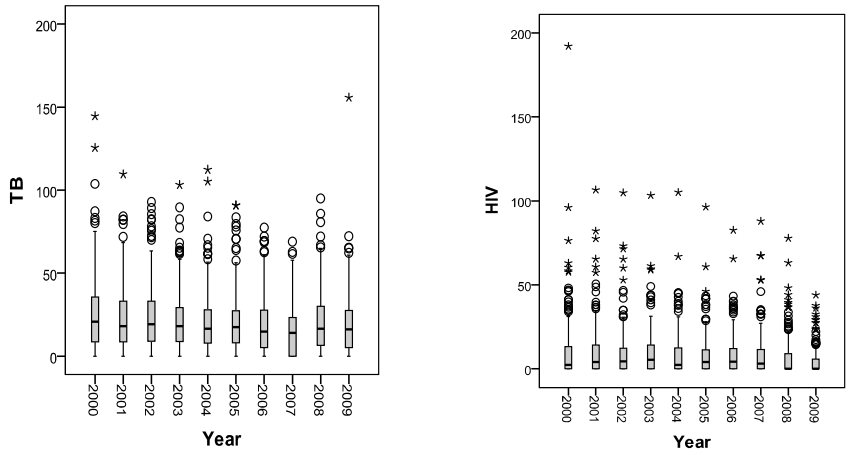


Figure 1: Boxplots of Municipalities TBIR and HIVIR, per year.

The presence of outliers, in most years, makes it difficult to interpret a global time trend for the incidence rates, regarding each disease (Figure 1).
As shown, from the minimum ("0" in all cases), the maximum and the median values, as well as from the boxplots, the data distribution was highly asymmetric for either TBIR or HIVIR, partly due to the strong presence of those outliers.

Table 3. Spearman Correlation Coefficient between TBIR and HIVIR (p<0.001).

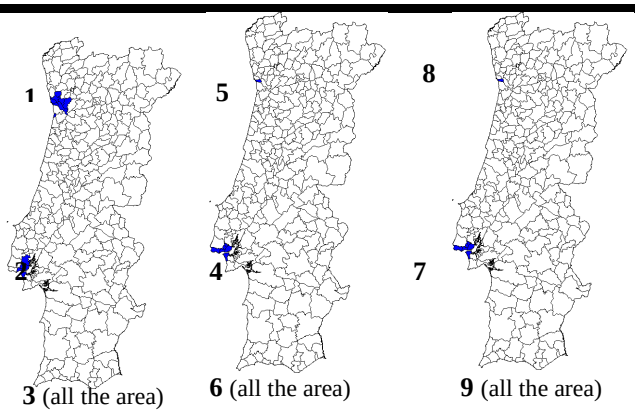
Global	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
0.279	0.294	0.342	0.308	0.341	0.204	0.280	0.313	0.281	0.213	0.202

The associations between TBIR and HIVIR, globally and per year, as expressed by Spearman correlation coefficients, seem stable but rather weak, though significant (Table 3). Correlation varied from 0.202 to 0.342 between 2000 and 2009, and it was 0.279 globally (p<0.001).

Table 4 & Figure 2. Results of TBIR(a) and HIVIR(b) spatiotemporal clustering analyses; (c) Multivariate Scan of TBIR and HIVIR (p<0.001).

Cl	Radius (km)	Time Frame	Observed/ Expected	R.R.
1	24543	2000-2009	6926/4345	1.77
2	13203	2000-2009	6563/4078	1.78
3	All	2000-2003	13561/11722	1.29
4	11591	2000-2009	6216/2467	3.31
5	0	2000-2009	2192/429	5.66
6	All	2000-2004	10692/9004	1.46
7	11591	2000-2009 2000-2009	6216/2467 6447/4029	3.31 1.77
8	0	2000-2009 2000-2009	2192/429 1668/701	5.66 2.46
9	All	2000-2004 2000-2004	10692/9004 16545/14705	1.46 1.28

Maps



a)TBIR b)HIVIR c)TBIR/HIVIR

For both diseases, Oporto and Lisbon Metropolitan Areas were identified as critical locals, based on spatiotemporal clustering analyses (Table 4 & Figure 2). Relative risks for Oporto and Lisbon were, respectively, 1.77 and 1.78 in TB study, and 5.66 and 3.31 in HIV case. Clusters 3, 6 and 9 were related to the whole area under study and only for the first years (until 2003 or 2004). This fact is concordant with the decreasing apparent trend already mentioned for both diseases.

4. Concluding remarks

The evidence of some matching between high incidence critical areas regarding both diseases is expected, in accordance with the scientific literature. Joint distributions of HIV/AIDS and Pulmonary Tuberculosis in space and time in Portugal Mainland were, in fact, not independent: very similar space-time critical areas were found, reinforcing the previous conviction that, in Portugal, HIV/AIDS may be faced as an explanatory, but not a major dimension for TB incidence. Oporto and Lisbon metropolitan areas were identified as important places for urgent Public Health interventions.

In order to improve the characterization of TB and HIV correlation, there is a need to: confirm that the detection rates are not interfering with results; explore the role of high outlier values as a source of joint variability; better understand the role of demographic and socio-cultural dimensions in the apparent associations; develop individuals-based studies, as a complement this ecological approach.

References

Anandaiah A., Dheda K., Keane J., Koziel H., Moore D.A., Patel N.R (2011) Novel developments in the epidemic of human immunodeficiency virus and tuberculosis coinfection, *American Journal of Respiratory and Critical Care Medicine*, **183**(8), 987-97.

Bhagyabati Devi S., Santa Naorem, Jeetenkumar Singh T., Birendra Singh Ksh, Lallan Prasad, Shanti Devi Th. (2005) HIV and TB Co-infection: A Study from RIMS Hospital, *Journal, Indian Academy of Clinical Medicine*, **6**(3), 220-3.

Couceiro L., Santana P., Nunes C (2011) Pulmonary tuberculosis and risk factors in Portugal: a spatial analysis. *The International Journal of Tuberculosis and Lung Disease* (*in press*).

de Colombani P., Banatvala N., Zaleskis R., Maher D. (2004) European framework to decrease the burden of TB/HIV, *European Respiratory Journal*, **24**(3), 493-501.

Kulldorff M. (1997) A spatial scan statistic, *Communications in Statistics: theory and methods*, **26**(6),1481–96.

Sanchez M.S., Lloyd-Smith J.O., Getz W.M. (2010) Monitoring Linked Epidemics: The case od tuberculosis and HIV, *Plos One* **5**(1):e8796

Spatial diffusion and temporal evolution of PCDD/Fs, PCBs and PAHs congener concentrations in the ambient air of Taranto: an analysis based on the duality diagram approach

Alessio Pollice
Dipartimento di Scienze Statistiche "Carlo Cecchi"
Università degli Studi di Bari "Aldo Moro"

Vittorio Esposito
Dipartimento di Taranto - Polo di Specializzazione Microinquinanti
ARPA Puglia

Abstract: Dioxins and dioxin-like compounds are byproducts of industrial processes, commonly regarded as highly toxic persistent organic pollutants. Polycyclic aromatic hydrocarbons occur in oil, coal, and tar deposits and are produced as byproducts of fuel burning, coke-making, and metal smelting. We propose an analysis of the spatial diffusion and temporal evolution of 46 congeners, based on monthly concentration data for the period October 2008 - December 2010 at three monitoring stations. Given the high dimensionality of the data, a descriptive strategy was adopted based on the duality diagram approach, a unifying framework including classical multivariate statistical methods that has become a valuable tool for combining data collected from different sources and using different methods.

Keywords: Air quality, Duality diagram, Multiple factor analysis

1 Introduction

Dioxins, dioxin-like compounds and polycyclic aromatic hydrocarbons are of concern because some compounds have been identified as carcinogenic, mutagenic, and teratogenic. The urban district of Taranto sits in close proximity to an industrial area where several large combustion plants are located, including an integrated cycle steel plant, an oil refinery, three waste incinerators, two power plants, and cement-works.

2 Materials and Methods

We propose an analysis of the spatial diffusion and the temporal evolution of 46 congeners split into five groups (PCDD, PCDF, PCB, LPAH, HPAH), based on monthly concentration data for the period October 2008 - December 2010 at three monitoring stations located within the industrial area (MA) and in a traffic/background area (AA and TA). Given the high dimensionality and multicollinearity of the available data, a descriptive strategy was adopted to obtain a synthesis of the spatio-temporal behavior and of the relationships between congeners. The duality diagram approach is a unifying framework including many classical multivariate statistical methods and less well-known recently developed tools for combining data collected from different sources and taking advantage of complex data types (Thiolouse, 2011). Within the K-table methods class, Multiple Factor Analysis (MFA) studies several groups of variables defined on the same set of observations and weights each group to achieve a joint representation of individuals and variables inducing a global representation of the groups of interest (Escoufier and Pagés, 1994). Weighting of variables groups is necessary to make the influence of each group comparable in a global analysis. MFA produces a display in which representations of the set of individuals associated to each group of variables are superposed. A global representation of groups of variables is obtained, in which each group is represented by the scalar product matrix it defines on the set of individuals. MFA search for factors which are common to several groups of variables is addressed by first setting up general variables, each one related to all the groups, and then searching for the canonical variable in each group for each variable. Each group defines a structure on the individuals set expressed by the shape of a cloud. A superposed representation which sets up the structure common to the different clouds is obtained in order to compare clouds one to another. A display in which each group is represented by one point allows to get a global comparison of groups.

The first step (interstructure) provides the coefficients of a special linear combination of the data tables, leading to an optimal summary called "compromise". The second step computes the PCA of this linear combination. The third step (intrastructure) is a projection of the rows and columns of each table into the multidimensional space of the compromise analysis. Functions implementing the methods used in this case-study are contained in the library ade4 of the statistical computation environment R (Dray et al., 2007).

3 Results

MFA is applied to the available data considering 15 groups of variables obtained crossing the 3 monitoring stations with the 5 groups of congeners. The first two principal components of the compromise account for 54% of the total inertia. With such a complex data structure this amount is considered sufficient for exploring the main features of the data. The behavior of the subsequent principal components

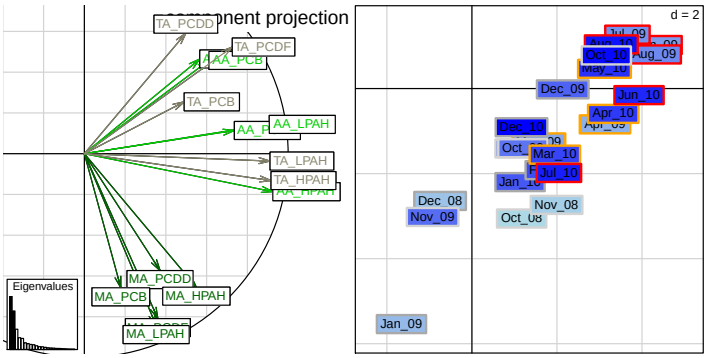


Figure 1: Left, projection of the principal components of each table on the compromise. Right, projection of the 23 observations on the compromise principal axes. Labels with darker background color for most recent time points are framed according to the season (red=Summer, orange=Spring, light grey=Autumn, dark grey=Winter).

(not reported) does not add any relevant information to the results obtained by the observation of the first two.

In Fig. 1, Left the industrial MA monitoring station appears to be quite separate from AA and TA (traffic/background). For the latter two, PAH's behave differently from all other congeners. A general seasonality can be seen in Fig. 1, Right with Spring/Summer observations in the top-right corner and Autumn/Winter on the opposite side.

Some more insights can be gained by looking at the biplots of each table in Fig. 2. For the MA monitoring station we observe a reduction in PCDD/F's and an increase od HPAH's. An overall reduction of the 46 congeners is registered at AA, while PAH's show a more evident decrease at TA. A stronger evidence of seasonal congener concentration decrease is found for PAH's at AA and TA.

4 Concluding remarks

Evidence of proximity to pollution sources is gained for the MA monitoring station given both the less pronounced seasonality, compared to AA and TA, and the contrasting trend of different pollutant groups, where PCDD/F show a general decrease while HPAH's are increasing over time. Industrial sources appear to be a major contributor to PAH pollution for MA compared to civil sources like traffic or domestic heating. While experiencing the impact of a comparable amount of road traffic, the MA station, where civil and industrial sources mix, has markedly different congener relationships compared to AA and TA.

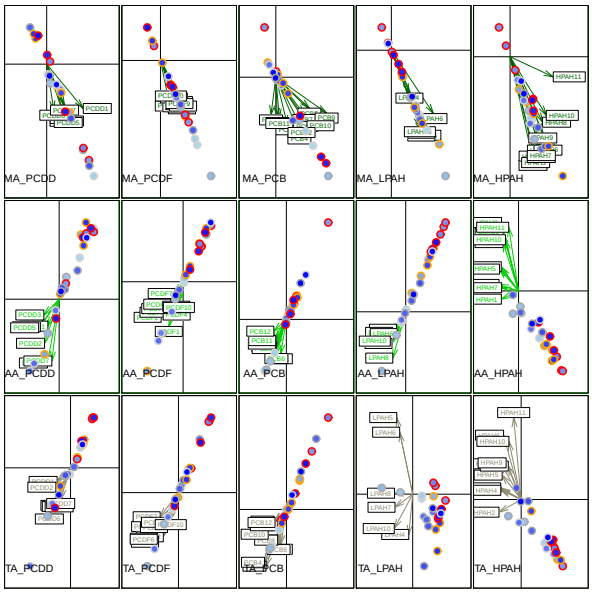


Figure 2: Biplots of the 15 tables (rows projected on the principal axes, columns projected on the principal components of the compromise). Dots with darker color for most recent time points are circled according to the season (red=Summer, orange=Spring, light grey=Autumn, dark grey=Winter).

References

Dray S., Dufour A.B., Chessel D. (2007) The ade4 Package - II: Two-table and K-table Methods, *R News*, 7 (2).

Escoufier B., Pagés J. (1994). Multiple factor analysis (AFMULT package). *Computational Statistics and Data Analysis*, 18, 121-140.

Thioulouse J. (2011) Simultaneous analysis of a sequence of paired ecological tables: a comparison of several methods, *Annals of Applied Statistics*, to appear.

Spatial disaggregation of pollutant concentration data [1]

Joanna Horabik, Zbigniew Nahorski
Systems Research Institute of Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland, Joanna.Horabik@ibspan.waw.pl

Abstract: The purpose of this study is to develop a method for allocating pollutant concentrations to finer spatial scales conditional on covariate information observable in a fine grid. Spatial dependence is modeled with the conditional autoregressive structure. The maximum likelihood approach to inference is employed, and the optimal predictors are developed to assess missing concentrations in a fine grid. The method is developed for a practical application of an output from the dispersion model CALPUFF run for Warsaw agglomeration.

Keywords: Air pollutant concentration, conditional autoregressive structure, spatial disaggregation

1 Introduction

Atmospheric dispersion models constitute a basic tool for air quality control. Further usage of output from dispersion models include, among others, health impact assessments. For improved risk assessments, it is often required to develop air quality data in a resolution higher than the one readily available from dispersion models.

Making inference on variables at points or grid cells different from those of the data is referred to as the change of support problem. Several approaches have been proposed to address the problem. The geostatistical solution for realignement from point to areal data is provided by block kriging (Gotway & Young 2002, Gelfand 2010). In the case that data are observed at areal units and inference is sought at a new level of spatial aggregation, areal weighting offers a straightforward approach. Some improved approaches with better covariate modeling were also proposed e.g. in Mugglin & Carlin 1998, and Mugglin *et al.* 2000.

In the following we present an approach for areal to areal data realignement, which accounts for a tendency toward spatial clustering, and is focused on application to air quality. The idea stems from the method proposed in Chow & Lin (1971) for time series, see also Polasek *et al.* (2010). Regarding an assumption on residual covariance structure, we apply the conditional autoregressive (CAR) specification. While the CAR structure is extensively used in epidemiology, it can be also applied for modeling air pollution over space (Kaiser *et al.* 2002, McMillan *et al.* 2010).

[1]The research of Joanna Horabik was supported by Ministry of Science and Higher Education under the Iuventus Plus project No. 0128/H03/2010/70.

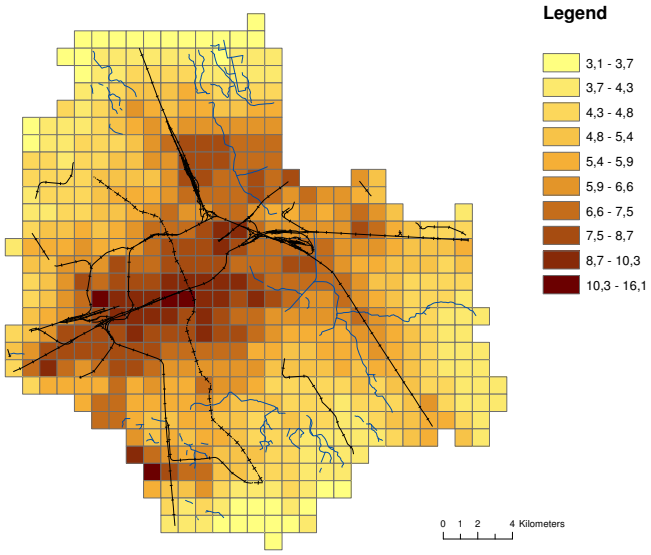


Figure 1: SO_2 concentration ($\mu g/m^3$) in a 1 km grid

2 Motivating data set

The study concerns air pollution concentrations (PM_{10}, NO_x and SO_2 among others) obtained from the dispersion model CALPUFF. A 1 km grid for Warsaw area comprises 563 grid cells. Health risk studies, conducted in parallel, motivated our search for the air pollution map in a 0.5 km resolution. The dispersion model output represents an average pollutant concentration over each 1 km grid cell. This value, multiplied by a cell area, reflects a pollutant level in a grid cell, and it constitutes the value to be disaggregated.

In addition, available covariate information characterizes transportation, area and point emission sources of the city in a 0.5 km grid.

3 The disaggregation framework

We begin with the model specification in a fine 0.5 km grid. Let Y_i denote a random variable associated with a missing value of pollutant, say SO_2, level y_i defined at each cell i, $i = 1, ..., n$ of a fine grid. Assume that random variables Y_i follow a Gaussian distribution with the mean μ_i and variance σ_Y^2, and given these values Y_i are independent. The values $\boldsymbol{\mu} = \{\mu_i\}_{i=1}^n$ represent the true process underlying SO_2 level, and the (missing) observations are related to this process through a measurement error of variance σ_Y^2. The model for the underlying SO_2 process is formulated as a sum of regression component with available covariates, and a spatially varying random effect. The applied CAR structure follows an assumption of similar random

effects in adjacent cells, and it is given through the specification of full conditional distribution functions

$$\mu_i|\mu_{j,j\neq i} \sim \mathcal{N}\left(\boldsymbol{x}_i^T\boldsymbol{\beta} + \rho\sum_{j\neq i}\frac{w_{ij}}{w_{i+}}\left(\mu_j - \boldsymbol{x}_j^T\boldsymbol{\beta}\right), \frac{\tau^2}{w_{i+}}\right),\ i,j = 1,...,n \tag{1}$$

where w_{ij} are the adjacency weights; w_{i+} is the number of neighbours of area i; $\boldsymbol{x}_i^T\boldsymbol{\beta}$ is a regression component with explanatory covariates for area i and a respective vector of regression coefficients, and τ^2 is a variance parameter. The joint distribution of the process $\boldsymbol{\mu}$ is (Cressie, 1993)

$$\boldsymbol{\mu} \sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}, \tau^2\left(\boldsymbol{D} - \rho\boldsymbol{W}\right)^{-1}\right), \tag{2}$$

where $\boldsymbol{X}$ is a design matrix with vectors $\boldsymbol{x}_i$; $\boldsymbol{D}$ is an $n\times n$ diagonal matrix with w_{i+} on the diagonal; and $\boldsymbol{W}$ is an $n\times n$ matrix with adjacency weights w_{ij}. Equivalently, we can write (2) as $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{N})$, with $\boldsymbol{N} = \tau^2\left(\boldsymbol{D} - \rho\boldsymbol{W}\right)^{-1}$.

The model for the CALPUFF output data observed in a 1 km grid is obtained by multiplication of $\boldsymbol{\mu}$ with an $N \times n$ *aggregation matrix* C, where N is a number of observations in a 1 km grid

$$\boldsymbol{C}\boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{\epsilon},\quad \boldsymbol{C}\boldsymbol{\epsilon} \sim \mathcal{N}_N\left(\boldsymbol{0}, \boldsymbol{C}\boldsymbol{N}\boldsymbol{C}^T\right). \tag{3}$$

The matrix C consists of 0's and 1's, indicating which cells have to be aligned together. The random variable $\boldsymbol{\lambda} = \boldsymbol{C}\boldsymbol{\mu}$ is treated as the mean process for variables $\boldsymbol{Z} = \{Z_i\}_{i=1}^N$ associated with observations $\boldsymbol{z} = \{z_i\}_{i=1}^N$ of the aggregated model

$$\boldsymbol{Z}|\boldsymbol{\lambda} \sim \mathcal{N}_N\left(\boldsymbol{\lambda}, \sigma_Z^2\boldsymbol{I}_N\right). \tag{4}$$

Also at this level, the underlying process $\boldsymbol{\lambda}$ is related to $\boldsymbol{Z}$ through a measurement error with variance σ_Z^2.

The parameters $\boldsymbol{\beta}, \sigma_Z^2, \tau^2$ and ρ are estimated with the maximum likelihood method based on the joint unconditional distribution

$$\boldsymbol{Z} \sim \mathcal{N}_N\left(\boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{M} + \boldsymbol{C}\boldsymbol{N}\boldsymbol{C}^T\right),$$

where $\boldsymbol{M} = \sigma_Z^2\boldsymbol{I}_N$. The analytical derivation is limited to the regression coefficients $\boldsymbol{\beta}$, and further maximisation of the profile log likelihood is performed numerically. The standard errors of estimators are calculated with the expected Fisher information matrix.

Regarding the missing values in a fine 0.5 km grid, the underlying SO_2 process is of our primary interest. The predictors optimal in terms of the minimum mean squared error are given by $E(\boldsymbol{\mu}|\boldsymbol{z})$. The joint distribution of $(\boldsymbol{\mu}, \boldsymbol{Z})$ is

$$\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{Z} \end{bmatrix} \sim \mathcal{N}_{n+N}\left(\begin{bmatrix} \boldsymbol{X}\boldsymbol{\beta} \\ \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{N} & \boldsymbol{N}\boldsymbol{C}^T \\ \boldsymbol{C}\boldsymbol{N} & \boldsymbol{M} + \boldsymbol{C}\boldsymbol{N}\boldsymbol{C}^T \end{bmatrix}\right). \tag{5}$$

The distribution (5) allows for full inference, yielding both the predictor and its error

$$\begin{aligned} \widehat{E(\boldsymbol{\mu}|\boldsymbol{z})} &= \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{N}}\boldsymbol{C}^T\left(\widehat{\boldsymbol{M}} + \boldsymbol{C}\widehat{\boldsymbol{N}}\boldsymbol{C}^T\right)^{-1}\left[\boldsymbol{z} - \boldsymbol{C}\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right] \\ \widehat{Var(\boldsymbol{\mu}|\boldsymbol{z})} &= \widehat{\boldsymbol{N}} - \widehat{\boldsymbol{N}}\boldsymbol{C}^T\left(\widehat{\boldsymbol{M}} + \boldsymbol{C}\widehat{\boldsymbol{N}}\boldsymbol{C}^T\right)^{-1}\boldsymbol{C}\widehat{\boldsymbol{N}}. \end{aligned}$$

Note that in the predictor $\widehat{E(\boldsymbol{\mu}|\boldsymbol{z})}$, a naive regression forecast is corrected with a residual on the aggregated level distributed over respective grid cells.

4 Concluding remarks

To conclude, the change of support problem in our study is addressed by defining the underlying air pollution process to be an aggregation for respective grid cells. The joint distribution (5) allows to view the approach in analogy to block kriging (Gelfand 2010, p.524).

The application part of the study is under development.

References

Chow G. C., Lin A. (1971) Best linear unbiased interpolation, distribution, and extrapolation of time series by related series, *The Review of Economics and Statistics*, 53, 372-375.

Cressie N.A.C. (1993) *Statistics for Spatial Data*, Wiley, New York.

Gelfand A.E. (2010) Misaligned Spatial Data: The Change of Support Problem, in: *Handbook of Spatial Statistics*, Gelfand A. E., Diggle P. J., Fuentes M., Guttorp P. (Eds.), Chapman & Hall/CRC, 517-539.

Gotway C.A., Young L.J. (2002) Combining incompatible spatial data, *Journal of the American Statistical Association*, 97, 632-648.

Kaiser M. S., Daniels M. J., Furakawa K., Dixon P. (2002) Analysis of particulate matter air pollution using Markov random field models of spatial dependence, *Environmetrics*, 13, 615-628.

McMillan N.J., Holland D.M., Morara M., Feng J. (2010) Combining numerical model output and particulate data using Bayesian space-time modeling, *Environmetrics*, 21, 48-65.

Mugglin A.S., Carlin B.P. (1998) Hierarchical modeling in geographical information systems: Population interpolation over incompatible zones, *Journal of Agricultural, Biological and Environmental Statistics*, 3,111-130.

Mugglin A.S., Carlin B.P., Gelfand A.E. (2000) Fully model-based approaches for spatially misaligned data, *Journal of the American Statistical Association*, 95, 877-887.

Polasek W., Llano C., Sellner R. (2010) Bayesian methods for completing data in spatial models, *Review of Economic Analysis*, 2, 194-214.

Spatial representativeness of an air quality monitoring station. Application to NO_2 in urban areas.

Maxime Beauchamp[a], Laure Malherbe[a], Laurent Létinois[a]
a : Institut National de l'Environnement Industriel et des Risques (INERIS), Direction des risques chroniques, Parc Technologique Alata, 60550 Verneuil-en-Halatte, France

Chantal de Fouquet[b]
b : Equipe géostatistique, centre de géosciences, Mines ParisTech, 35 rue Saint Honoré, 77305 Fontainebleau, France

Abstract: The present study aims at setting up a geostatistical methodology that could be implemented in an operational context to assess the spatial representativeness of a measurement station. In the proposed definition, a point is considered as belonging to the area of representativeness of a station if its concentration differs from the station measurement by less than a given threshold. Additional criteria related to distance or environmental characteristics may also be introduced.

Concentrations are first estimated at each point of the domain applying kriging techniques to passive sampling data obtained from measurement surveys. The standard deviation of the estimation error is then used, making a hypothesis on the error distribution, to select the points, at a fixed risk, where the difference of concentration with respect to the station is below the threshold.

The methodology is then applied to NO_2 experimental datasets for different French cities.

Keywords: geostatistics; kriging; spatial representativeness; nitrogen dioxide (NO_2).

1. Introduction

Local agencies in charge of air quality monitoring are concerned with assessing the geographical areas in which concentrations may be assumed similar to those measured by monitoring stations.

Spatial representativeness of a monitoring site is a recurrent notion that appears in European regulatory requirements on air quality but has not been precisely defined so far. A definition will be proposed and its practical implementation will lead to the production of maps to characterize areas represented by the stations.

Application of the method for the background pollution [1] will be presented and some issues concerning the consideration of a traffic-related pollution model will be discussed.

2. Materials and Methods

First, an estimation of the NO_2 annual average of the background pollution is provided at each point of the domain applying kriging techniques to passive sampling surveys data. High resolution auxiliary variables, like the NOx emissions density in a 2km radius are also used as external drift.

A first approach to define the area of representativeness of a monitoring station S_0 located in x_0 is to consider all the sites where the concentrations are sufficiently close to the station measurement, which implies the introduction of a threshold notion [2][3][4]:

$$|Z(x) - Z(x_0)| < \delta \qquad (E.1)$$

Let's consider the estimation error of the pollution $\varepsilon(x) = Z(x) - Z^*(x)$. We don't take the measurement error at the station into account.

$$|Z(x) - Z(x_0)| < \delta$$

$$\Leftrightarrow \; |Z^*(x) + \varepsilon(x) - Z(x_0)| < \delta \qquad (E.2)$$

A sufficient condition for (E.2) is:

$$|\varepsilon(x)| < \delta - |Z^*(x) - Z(x_0)| \qquad (E.3)$$

We introduce the statistical risk η that the concentration of a point considered in the area of representativeness of S_0 differs from the station measurement by more than the given threshold δ:

$$\mathbb{P}[|\varepsilon(x)| \geq \delta - |Z^*(x) - Z(x_0)|] < \eta \qquad (E.4)$$

Then, making a Gaussian hypothesis on the error distribution, the standard deviation of the estimation error is used to select the points in the area of representativeness:

$$|Z^*(x) - Z(x_0)| < \delta - \sigma_K(x) * q_{1-\frac{\eta}{2}} \qquad (E.5)$$

In this approach, a point can be considered as belonging to several areas of representativeness. So, additional criteria related to distance, minimal deviation of concentration, or environmental features are introduced to make a point belong to a unique station.

Local scale also enables to estimate concentrations taking traffic-related pollution into account: distance to the road, traffic-related NOx emissions, or road traffic informations can be considered to develop and improve a model.

3. Results

To illustrate the results of the methodology, passive sampling data provided by a survey carried out in the French city of Montpellier in 2007 are used.

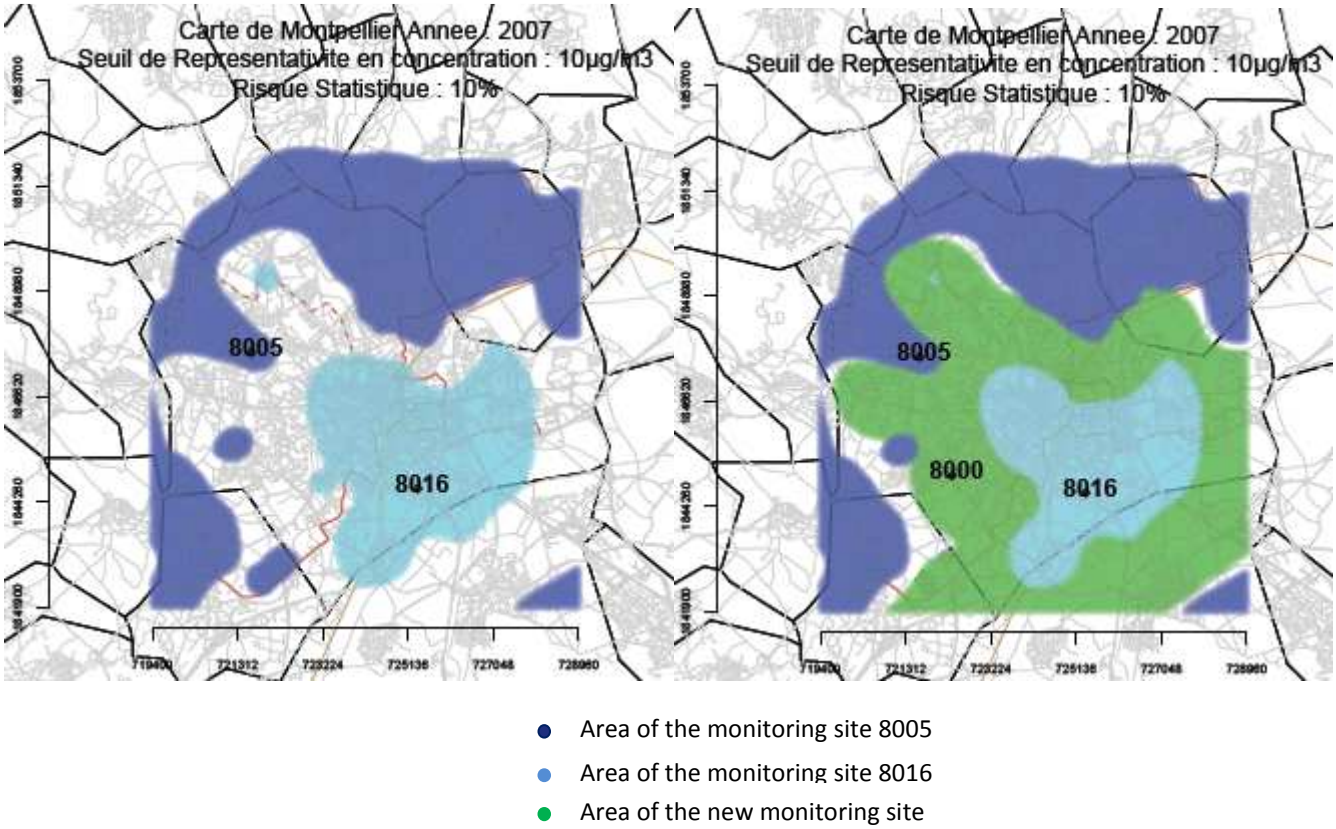


Figure 1: Areas of representativeness of the background monitoring sites for the French city of Montpellier in 2007, for a threshold δ of 10µg/m^3 and a risk η fixed at 10%

Figure 1 shows the application of the method on the background pollution for a threshold δ of 10µg/m^3 and a statistical risk η fixed at 10%. Two areas of representativeness can be obtained: a first one for the downtown pollution and a second one for the suburb pollution.

Results can be helpful in providing some recommendations for setting up new fixed monitoring sites. In this case, sampling passive data can be used to find an appropriate site where the concentration of NO_2 is the most representative of the missing information.

4. Concluding remarks

Application of the method for background pollution using analyzed data of NO_2 annual concentrations produced on national scale shows its sensitivity to the criterion selected to remove intersections between representativeness areas. Stability in time of the areas is also related to variations of concentrations on the domain.

This study underlines the difficulty to set up a reliable traffic-related pollution model and the influence of the passive sampling data location on the quality of the model.

The way of taking account of the error of this traffic-related pollution model could also be discussed in future studies: the introduction of a Gaussian hypothesis as well as the

results of Chilès and Delfiner under a continuous and unimodal distribution error [5] are envisaged.

References

[1] ADEME (2002). Classification et critères d'implantation des stations de surveillance de la qualité de l'air.
[2] Bobbia M., Cori A., De Fouquet C. (2008). Représentativité spatiale d'une station de mesure de la pollution atmosphérique. Pollution Atmosphérique N°197.

[3] Cori A. (2005). Représentativité spatiale des stations de mesure de la concentration moyenne annuelle en NO2. Rapport de stage. École des Mines de Paris.

[4] Cardenas G. et Malherbe L. (2007). Représentativité des stations de mesure du réseau national de surveillance de la qualité de l'air. Application des méthodes géostatistiques à l'évaluation de la représentativité spatiale des stations de mesure de NO2 et O3.
Report available at www.lcsqa.org.

[5] Chilès J-P and Delfiner P. (1999). Geostatistics: Modeling Spatial Uncertainty. Wiley Series in Probability and Mathematical Statistics, 695 p.

Statistical investigations on PAH concentrations at industrial sampling site

Martino Amodio, Eleonora Andriani, Paolo R. Dambruso, Gianlugi de Gennaro*, Annamaria Demarinis Loiotile, Alessia Di Gilio, Livia Trizio
Chemistry Department, University of Bari, giangi@chimica.uniba.it

Giorgio Assennato, Candida Colucci, Vittorio Esposito, Roberto Giua, Micaela Menegotto, Maria Spartera
Apulia Region Environmental Protection Agency (ARPA Puglia)

Abstract: Human exposure to combustion emissions including the associated airborne fine particles and mutagenic constituents have been studied in populations in different countries. PAH compounds are generated by combustion of organic matter in mobile sources as well as in stationary sources; they play a major role in defining the overall toxicity of atmospheric particulate matter (PM) although they are negligible in the total mass of the PM. The aim of this work was to apply statistical investigations on PAH concentrations measured at industrial sampling site in Taranto (Apulia Region, South of Italy) from May 2009 to May 2010. These data, related to gaseous pollutants at different meteorological conditions, allowed to determine the relationships between industrial emissions and ambient concentrations at receptor site.

Keywords: PAHs, industrial site, emission sources

1. Introduction

Several epidemiological studies suggested the relevant role of ambient Particulate Matter (PM) in contributing to a range of health effects: the increased risk of death has been associated to the exposure to high PM concentrations, especially to the finer particles (Nadadur et al., 2009). In particular, it was found that the finer particles (PM2.5) can transport the pollutants deeply into the lung and cause many kind of reactions which include oxidative stress, local pulmonary and systemic inflammatory responses (Englert, 2004; Forbes et al, 2009). Particulate Matter consists of major components representing the main part of the total mass of particles and trace components usually representing less than 1% of total particle mass. Among the PM trace components, Polycyclic Aromatic Hydrocarbons (PAHs) constitute a major class of environmental pollutants. Many PAHs, particularly the larger five- and six-ring compounds that can be metabolized to diol epoxides, are mutagens and carcinogens (Binkova et al. 2007; de Kok et al. 2006). The primary source of PAH compounds in air pollution is from combustion of fossil fuels (e.g., coal, oil, gasoline and diesel fuel), vegetative matter (e.g., wood, tobacco, paper products, and biomass) and synthetic chemicals (e.g., from plastics and other chemical products in incinerated municipal, hospital and hazardous wastes). Once released in atmosphere, PAHs are subjected to

several atmospheric processes; heterogeneous reactions (photo-oxidations) and gas-particle partitioning are the main transformations processes of PAHs. These processes are dependent on the different meteorological conditions. The aim of this work was to assess the effects of emission sources on particle-bound PAH concentrations determined at the sampling site in Taranto from May 2009 to May 2010. These data were linked to meteorological conditions and gaseous pollutants measured at the monitoring station (NO_x, CO, BTX). Finally, Principal Component Analysis (PCA) was applied to the dataset in order to provide information on the most relevant emission sources located in the area under investigation.

2. Materials and Methods

The sampling site is located in Taranto (Apulia Region, South of Italy), close to the industrial area (Via Machiavelli, Tamburi district). The sampling station is a customised monitoring unit containing a range of real-time instruments for particulate matter and gaseous pollutants. It includes a PM_{10} analyzer (SWAM Monitor), a nitrogen oxides ($NO_{x)}$ analyzer (API200A), a carbon monoxide (CO) analyzer (API300), a benzene-toluene-xylene (BTX analyzer) (Syntech Spectras) and a particle-bound PAHs analyzer (EcoChem PAS 2000). Meteorological data (wind direction, wind speed, air temperature, rainfall, barometric pressure and solar irradiation) are continuously recorded by an automated weather station.
Principal Component Analysis (PCA) was applied to the pollutant concentrations determined at the sampling site in order to obtain information on the characteristics of the most relevant emission sources located in the area.

3. Results

PAHs data collected from May 2009 to May 2010 were related to meteorological conditions, in particular to wind speed and direction determined at the sampling site. The higher PAHs concentrations were observed for air masses coming from North-East and North-West (Sector I, 0° - 60° (North) e 300°-360° (North)) directions because of the presence of the industrial area. As shown by previous studies at the same sampling site (Amodio et al., 2009), the days characterized by gust from North or by calm wind conditions coincided with the maxima PAHs values. Therefore, the data collected in the investigated period were related to wind speed for the higher concentration direction (Sector I, 0° - 60° (North) e 300°-360° (North)) and for the other ones (Sector II, 60° - 180° (North); Sector III, 180° - 300° (North)). As shown in Figure 1, the higher PAHs concentrations were found for air masses coming from the North (Sector II), even when high wind speed is determined at the sampling site. However, as concern Sector II, lower PAHs concentrations than those previously observed can be measured at receptor site when air masses come from East direction. It was also found that the most significant values for PAHs were observed when calm wind conditions occurred, maybe due to low dispersion capacity of the pollutants in the atmosphere. Finally, the analysis of data collected in Sector III (180° - 300° (North)) (data not shown) showed the similar trend of those in Figure 2; some additional high events of pollutant concentrations were

determined for air masses come from West - North West direction, as observed for 'high polluted' direction.

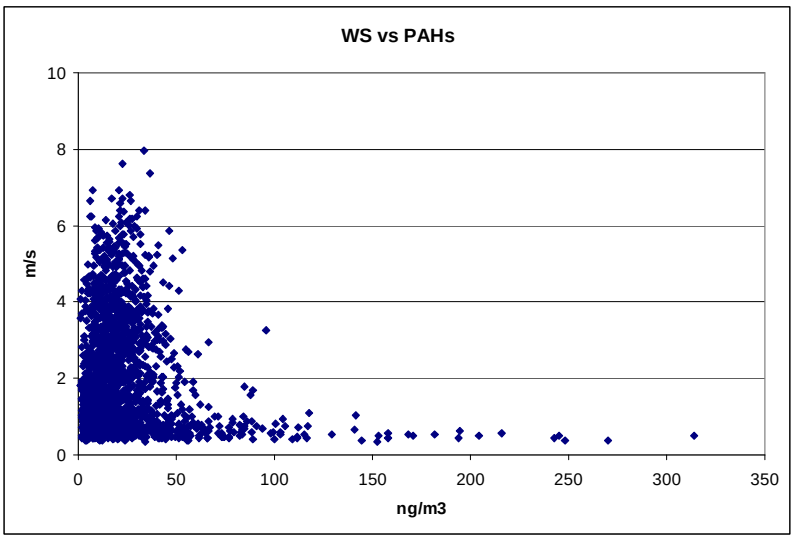


Figure 1: Sector I, wind speed (m/s) versus PAHs concentrations (ng/m^3)

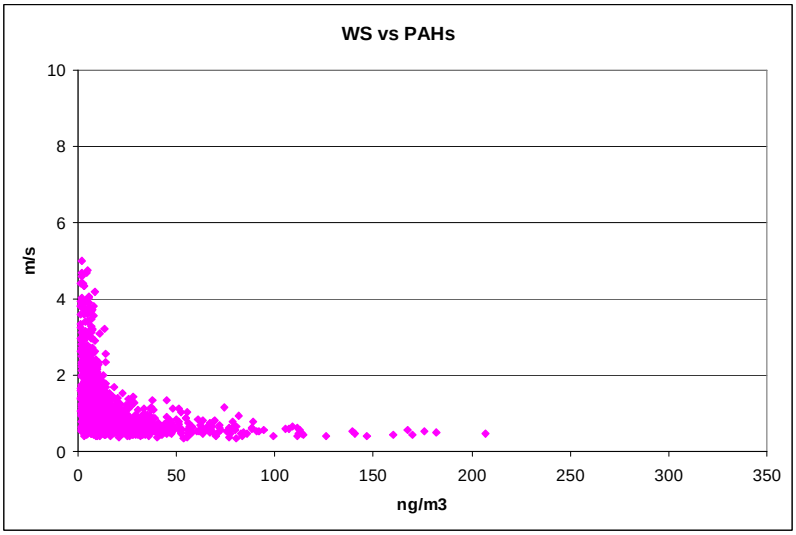


Figure 2: Sector II, wind speed (m/s) versus PAHs concentrations (ng/m^3)

Principal Component Analysis with Varimax normalized rotation was applied on the data matrix of hourly mean concentrations of nitrogen oxides (NO_x), carbon monoxide (CO), benzene (B), toluene (T) and total PAHs. Since the variables were characterized by different orders of magnitude, PCA was applied to normalized data matrix. Loadings and percentage of explained variance obtained for each of the components are shown in Table 1; only variables with factor loadings greater than 0.3 are shown. Two PCs, explaining up to 80% of the total variance of data, were evaluated. The PC1, which explained the most of the variance of the data, is characterized by high loadings for all the considered parameters, except of benzene (PC2). It was explained with the closeness of the monitoring station to the industrial area, that significantly affect the pollutant

concentrations measured at the receptor. In fact, the same results in PCs, loading and explained variable (data not shown) were obtained when the PCA was performed to the dataset containing the samples collected in Sector I.

	Loadings	
	1	2
NOx	0.93	
CO	0.79	0.35
BENZENE		0.97
TOLUENE	0.66	
PAHs	0.94	
% Var	56.97	23.39

Table 1: Loadings, eigenvalues and percentage of explained variance obtained in PCA on data collected in Taranto sampling site

4. Concluding remarks

This work was performed on data collected at Taranto sampling site (Via Machiavelli, Tamburi district) from May 2009 to May 2010. It allowed to highlight the relevance of the industrial area closed to the receptor site, that caused high pollution events when the air masses flow from the North. The same results were obtained by taking into account Principal Component Analysis performed on the dataset.

References

Amodio M., Andriani E., Caselli M., Dambruoso P.R., Daresta B.E., de Gennaro G., Ielpo P., Placentino C.M., Tutino M. (2009) Characterization of particulate matter in the Apulia Region (South of Italy): features and critical episodes, *Journal of Atmospheric Chemistry*, 63(3), 203-220.

Binkova B., Topinka J., Srama R.J., Sevastyanova O., Novakova Z., Schmuczerova J., Kalina I., Popov T., Farmer P.B. (2007) In vitro genotoxicity of PAH mixtures and organic extract from urban air particles Part I: Acellular assay, *Mutation Research*, 620, 114-122.

de Kok .M.C.M., Driece .A.L., Hogervorst .G.F., Briede J.J. (2006) Toxicological assessment of ambient and traffic-related particulate matter: A review of recent studies. *Mutation Research*, 613, 103-122.

Englert N. (2004) Fine particles and human healt - a review of epidemiological studies, *Toxicology Letters*, 149(1-3), 235-242.

Forbes L., Patel M.D., Rudnicka A.R., Cook D.G., Bush T., Stedman J.R., Strachan D.P., Anderson H.R. (2009) Chronic exposure to outdoor air pollution and diagnosed cardiovascular disease: meta-analysis of three large cross-sectional surveys, *Environmental Health*, 8(30).

Nadadur S.S, Miller C.A., Hopke P.K, Gordon T., Vedal S., Vandenberg J.J., Costak D.L. (2007) The complexities of air pollution regulation: the need for an integrated research and regulatory perspective, *The Journal of Toxicological Sciences*, 100, 318-327.

Tapering spatio temporal models [1]

Alessandro Fassó, Francesco Finazzi, Moreno Bevilacqua
DIIMM, University of Bergamo, alessandro.fasso@unibg.it

Abstract: We consider regression models for multivariate spatio temporal data. We view the data as a time series of spatial processes and work in the setting of dynamic models. In order to add flexibility we consider regression models with spatio temporal varying coefficients. Spatial dependence among the different measurements is attained considering the linear model of coregionalization. Since spatio-temporal data are tipically of large dimension we propose to perform estimation both through maximum likelihood by means of the EM algorithm and a modified version of it exploiting the covariance tapering likelihood function.

Keywords: Air quality assessment, Covariance tapering, EM algorithm,r maximum likelihood estimation.

1 Introduction

The increasing availability of datasets on multivariate spatio-temporal data parallels the need for statistical models which are flexible enough for covering the underlying complexity and can be estimated by means of well founded inferential techniques. The dynamic coregionalization model, recently proposed by Fassó and Finazzi (2011a), has these advantages as it allows modelling of complex multivariate spatio-temporal dynamics and performing maximum likelihood parameter estimation by means of the EM algorithm.

Due to the advancement of technology, massive amounts of data are often observed at a large number of spatial locations in environmental sciences. For this reason recent literature focused on geostatistical analysis of large multivariate spatio-temporal datasets. See for instance Bevilacqua *et al.* (2011) and Cressie and Johannesson (2008). This is because spatial problems with modern data often overwhelm traditional implementations of spatial statistics, such as maximum likelihood estimation. In this paper, in order to estimate multivariate regression spatio temporal models for EU air quality assessment, we consider an approximation of the estimation method proposed by Fassó and Finazzi (2011a) by considering the covariance tapering approach. The key idea is that the use of covariance tapering allows to manage large multivariate spatio temporal data.

[1]This research is part of Project EN17, Methods for the integration of different renewable energy sources and impact monitoring with satellite data", funded by Lombardy Region under Frame Agreement 2009.

2 Dynamic coregionalization model with varying coefficients

We consider the following observation equation for the multivariate spatio-temporal random process $Y(s,t) = (Y_1(s,t), \dots Y_q(s,t))'$ at time $t = 1, \dots, T$ and site $s \in D \subset R^2$:

$$Y(s,t) = X_1(s,t)\beta + X_2(s,t)K_2Z(t) + X_3(s,t)K_3W(s,t) + \varepsilon(s,t), \qquad (1)$$

where X_1, X_2, X_3 are matrices of known covariates and K_2, K_3 are matrices of constants . $Z(t)$ is a $p-$variate Markovian component, $W(s,t)$ is a $r-$variate Gaussian random field and $\varepsilon(s,t)$ is a $q-$variate Gaussian white noise in space and time.

The $p-$dimensional latent temporal state $Z(t)$ has the Markovian dynamics $Z(t) = GZ(t-1) + \eta(t)$, with G a stable transition matrix and $\eta \sim N(0, \Sigma_\eta)$. The $k-$dimensional Gaussian random field is described by coregionalization model of c components

$$W(s,t) = \sum_{j=1}^{c} W^{[j]}(s,t)$$

where each $W^{[j]}(s,t)$, for fixed t, is a latent zero-mean Gaussian process with covariance and cross-covariance matrix function $\Gamma^{[j]} = cov\left(W_i^{[j]}(s,t), W_{i'}^{[j]}(s',t)\right) = V_j\rho^{[j]}\left(h, \overline{\theta}^{[j]}\right)$, $1 \le i, i' \le r$, $1 \le j \le c$. Each V_j is a coefficents matrix and each $\rho^{[j]}$ is a valid correlation function and $h = \|s - s'\|$ is the Euclidean distance between s and s'. All spatial processes above are purely spatial processes in the sense that are uncorrelated over different time points. Finally, $\varepsilon_i(s,t) \sim N(0, \sigma_{\varepsilon,i})$, $i = 1, \dots q$ is the measurement error which is white-noise in space and time. The parameter set to be estimated is $\Psi = (\beta, \sigma_\varepsilon; G, \Sigma_\eta; \theta; V) = (\Psi_Y, \Psi_Z, \Psi_W)$ where $\beta = (\beta_1, ..., \beta_q)'$, $\sigma_\varepsilon = (\sigma_{\varepsilon,1}, ..., \sigma_{\varepsilon,q})'$, $\theta = (\theta_1, ..., \theta_c)'$ and $V = (V_1, ..., V_c)'$.

3 Estimation method

At each time t, each $Y_i(s,t)$ is observed at n_i sites $S_i = (s_{i,1}, ..., s_{i,n_i})$. The sets in $S = (S_i, ..., S_q)$ are not constrained and can be disjoint. The observed vector at time t is then $Y(S,t) = (y_1(S_1,t), ..., y_q(S_q,t))'$ a vector of dimension $N = \sum_{i=1}^{q} n_i$.

Due to the Markovian assumption and to the space-time separability property of the model, and setting $Y = (Y(S,1), \dots, Y(S,T))'$, $Z = (Z_0, Z_1, ..., Z_T)'$, with $Z_t = Z(t)$ and $W^{[j]} = \left(W_1^{[j]}, ..., W_T^{[j]}\right)'$, $j = 1, .., c$, and $W = \left(W^{(1)}, ..., W^{(c)}\right)'$, the complete-data log-likelihood function $L(\Psi; Y, Z, W)$ takes the nice additive form:

$$l(\Psi; Y, Z, W) = l(\Psi_Y; Y \mid Z, W) + l(\Psi_Z; Z) + \sum_{j=1}^{c}\sum_{t=1}^{T} l\left(\Psi_W; W_t^{[j]}\right)$$

where

$$l\left(\Psi_Z; Z\right) = l\left(\Psi_Z; Z_0\right) + \sum_{t=1}^{T} l\left(\Psi_Z; Z_t \mid Z_{t-1}\right)$$

The involved distributions are all of Gaussian type. Specifically for the latent variable

$$\begin{aligned} Z_0 &\sim N_p\left(\mu_0, \Sigma_0\right) \\ \left(Z_t \mid Z_{t-1}\right) &\sim N_p\left(GZ_{t-1}, \Sigma_\eta\right) \\ W_t^{[j]} &\sim N_N\left(0, \Sigma^{[j]}\right),\ 1 \leq j \leq c \end{aligned}$$

Note that $\Sigma^{[j]} = \Sigma^{[j]}(\Psi_W)$ is a matrix of dimension $R \times R$. with $R = \sum_{i=1}^{r} n_i$

Estimation can be performed adapting the EM- algorithm as proposed in Fassó and Finazzi (2011a) and modified by Fassó and Finazzi (2011b).

Here we propose a modification of this alghorithm to take into account the problem of large dataset. Specifically the modification is based on the covariance tapering likelihood idea Kaufmann *et al.* (2008) that is we consider the tapered complete data likelihood:

$$l_{TAP}\left(\Psi; Y, Z, W\right) = l\left(\Psi_Y; Y \mid Z, W\right) + l\left(\Psi_Z; Z\right) + \sum_{j=1}^{c}\sum_{t=1}^{T} l_{TAP}\left(\Psi_W; W_t^{[j]}\right)$$

where $l_{TAP}\left(\Psi_W; W_t^{[j]}\right)$ is defined as:

$$l_{\mathrm{TAP}}(\Psi_W; W_t^{[j]}) = -\frac{1}{2}\log|\Sigma^{[j]} \circ T(d)| - \frac{1}{2} W_t^{[j]\prime}([\Sigma^{[j]} \circ T(d)]^{-1} \circ T(d)) W_t^{[j]} \qquad (2)$$

This is the multivariate version of the tapering likelihood proposed by Kaufman *et al.* (2008). In their approach for the univariate case, certain elements of the covariance matrix are set to zero multiplying it element by element by a correlation matrix coming from a compactly supported isotropic correlation function.

Here $T(d)$ is a sparse cross-correlation matrix coming from a valid model of matrix valued correlation function with compact support and $\circ$ is the Schur product. A simple model for the isotopic case is the following: let $\rho(h, d)$ a compact support correlation function (one of the Wendland (1995) class for instance) and let B a $r \times r$ positive definite matrix of coefficients, then $\rho(h, d)B$ is a $r \times r$ valid model of matrix valued correlation function with compact support. The associated matrix is $T(d) = B \otimes H(d)$ where $H(d) = \{\rho(||s_i - s_j||, d)\}_{i,j=1}^{n}$. The 'tapered' matrix $\Sigma^{[j]} \circ T(d)$ is still positive definite and sparse matrix algorithms can be used to evaluate an approximated likelihood efficiently. The intuition behind this approach is that correlations between pairs of distant sampling locations are often nearly zero, so little information is lost in taking them to be independent.

References

Bevilacqua, M., Gaetan, C., Porcu, E.,Mateu, J.. (2011), Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. *Journal of the American Statistical Association*, Accepted for publication.

Cressie, N., and Johannesson, G. (2008), Fixed rank kriging for very large spatial data sets, *Journal of the Royal Statistical Society*, Series B, 70, 209-226..

Fassó A., Finazzi F., (2011a) Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics* Accepted for publication.

Fassó A., Finazzi F. (2011b) Spatio temporal models with varying coefficients for European air quality assessment. Proceedings of Spatial 2, Spatial Data Methods for Environmental and Ecological Processes 2nd Edition, Foggia, Sept. 1-2, 2011.

Kaufman, C., Schervish, M., and Nychka, D. (2008), Covariance tapering for likelihood- based estimation in large spatial datasets, *Journal of the American Statistical Association*, 103, 1556-1569.

Wendland, H. (1995), Piecewise Polynomial, Positive Definite and Compactly Supported Radial Functions of Minimal Degree, *Advances in Computational Mathematics*, 4, 389-396.

A METHODOLOGY FOR EVALUATING THE TEMPORAL STABILITY OF SPATIAL PATTERNS OF VINEYARD VARIATION

Gambella F.[(1)]; Dau R.[(1)]; Paschino F.[(1)]
[(1)]University of Sassari, Department of Agricultural Engineering, Viale Italia 39, 07100 Sassari, Italy.

Castrignanò A.[(2)]; De Benedetto D. [(2)];
[(2)]Consiglio per la Ricerca e la sperimentazione in Agricoltura (CRA). Unità di ricerca per i sistemi colturali degli ambienti caldo-aridi (Bari) Via Celso Ulpiani 5, 70125 - Bari-Italy. annamaria.castrignano@entecra.it

Abstract: Vineyards vary substantially in the quantity and quality of grapes they produce. The study was undertaken in a commercial "Semidano" vineyard block (0.6 ha) in the municipality of Mogoro (Sardinia isle, Italy) during the vintages of 2008, 2009 and 2010. A total of 106 plants were sampled and georeferenced. To assess the joint spatial and temporal variation of the vine properties, a multivariate geostatistics technique was applied, called factor cokriging, which aims at decomposing the overall variance in a restricted number of regionalised scale-dependent factors. The thematic maps of the vineyard properties and the ones of the factors show a large variability on both space and time. All the measurements of spatial agreement reveal a lack of temporal stability of the variation patterns over the years.

Keywords: precision viticulture, geostatistical analysis, temporal stability

1.Introduction

Precision Viticulture (PV) is an application of new Information Technologies (IT) used to maximize grape production efficiency and quality while minimizing environmental impact and risk. It is actually dependent on the existence of spatial variability in either product quantity or quality or both. Some variables may also be temporally variable, but have stable spatial patterns or show little temporal stability. The most compelling argument for the adoption of PV is the accurate assessment of variability that has been observed in vegetative growth, yield and grape quality over the past few years (Bramley, 2004). The objectives of PV will differ, depending on the market requirements for wine, and the use of selective harvesting might be utilized to optimize quality (Bramley et al., 2003). Thus winemakers need to produce grapes that maintain certified characteristics of good quality over the years and to reduce the interventions by adopting site-specific techniques. Grape quality within the zones characterized by different vegetative vigor is tested by using a stratified sampling and the results are used to formulate differential harvest strategies (Bramley et al. 2005). When maps are delivered, farmers receive a large amount of data which has to be analyzed rapidly. This means that the decision, as to whether or not it is appropriate to apply site-specific management (SSM), has to be made in a few days. This step is even more critical in viticulture when the information is delivered and analyzed at the cooperative level. The

primary technological advance that makes precision agriculture feasible is the yield map, which enables the farmer to estimate crop yields for sections as small as a few square meters and to display the collection of these estimates in color-coded maps. Growers can identify high- and low-yielding regions of the field and precisely quantify the differences between them. To produce accurate maps of yield and grape quality, the use of geostatistics may be much valued but it is also needed to introduce a methodology for evaluating the temporal stability of spatial patterns in the vineyard. The objectives of this work were to delineate homogeneous zones within a vineyard and test their stability over the years.

2. Materials and Methods

This study was undertaken in a commercial "Semidano" vineyard block (0.6 ha) in the municipality of Mogoro (Sardinia, Italy) during the years 2008, 2009 and 2010. The plants were harvested at 106 locations georeferenced by using a DGPS device (GRS1, TOPCON), and the number of bunches, the average bunch weight and the total production per plant were determined for each year (9 variables). The Babo° degree of the grapes was calculated by rifractometric method using an optical refractometer (MR 210, Greensis) and the total acidity (tartaric acid) was expressed in ml of NaOH 0,1 N used for tritated 7.5 ml of must solution. The measurements were done only in 2010. The multivariate data set then included eleven variables.

To assess the joint spatial and temporal variation of the vineyard properties a multivariate geostatistical technique was applied, called factor cokriging analysis (FCKA), which aims at decomposing the overall variance in a restricted number of regionalised scale-dependent factors. The theory underlying FCKA was described by Castrignanò et al., 2000 and Wackernagel, 2003.

To produce the maps of the variables, MultiGaussian approach was used which requires a prior Gaussian transformation of the initial attribute into a Gaussian-shaped variable (Wackernagel, 2003, pp. 238-249).

The transformed data were then submitted to geostatistical analyses and the estimates were back-transformed to the raw data to produce thematic maps.

The three basic steps of FCKA are the following:

1) modelling the coregionalization of the set of variables, using the so called Linear Model of Coregionalization (LMC) and interpolating the variables at the nodes of a 1 by 1m-cell grid by cokriging;

2) analysing the correlation structure between the variables, by applying Principal Component Analysis (PCA) at each spatial scale;

3) cokriging specific factors at the characteristic scales and mapping them.

Contingency table and Cohen's kappa statistic were used to evaluate the stability of the spatial patterns of variation over the three-year period of study.

3. Results

The exploratory analysis (results not shown) of measured variables in the different years revealed considerable variation both across the vineyard and over time. The spatial variation was attested by the high values of CVs, whereas substantial temporal variability between the vintages was observed, as tested by the mean and maximum values, which varied significantly among the different dates for each type of variable.

A LMC was fitted to the Gaussian transformed data of all variables, including two basic structures: a nugget effect and an isotropic spherical model with range=18 m. The overall spatial and temporal variance was split into two approximately equal components, the not-spatially correlated component (nugget effect) and the short-range component. The previous results show that the most variation within the vineyard occurred at very short distance.

The spatial maps of the eleven variables, obtained by cokriging, show a large variability on both space and time making difficult to disclose some distinct spatial patterns (only the maps of grape mass per plant were reported in Figure 1). Nevertheless, there is a wide central area mostly characterised by the highest values of number of bunches, average bunch weight and total production per plant in all the three years and by the minimum values of acidity and Babo degrees of grapes, in 2010 characterized by a general increase in erraticity.

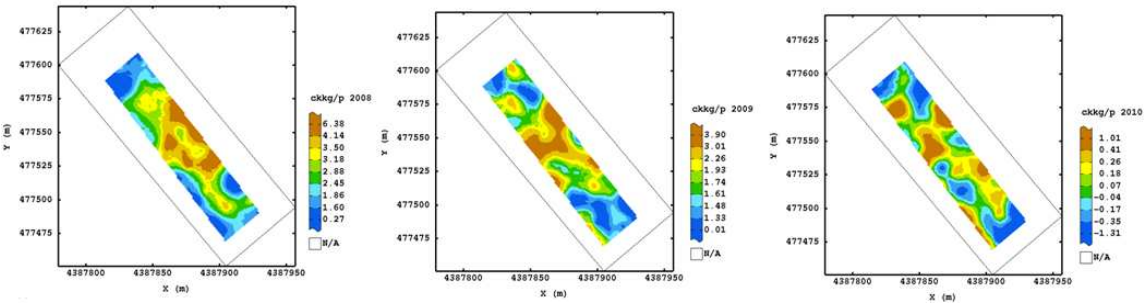


Figure 1: Thematic maps of grape mass per plant for the three years.

In PCA applied to 18m-range structure we retained the eigenvectors (factors) producing eigenvalues greater than one. Therefore, we focused on the first three factors, which accounted for 36%, 30% and 20%, respectively, of the total variation at the corresponding spatial scale. The positive loading values for the three factors (data not shown) indicated the variables recorded in 2009, 2008 and 2010 as the ones most influencing the first, second and third factor, respectively. These factors can then be assumed as indicators of grape production in 2009 and 2008 and also of grape quality in 2010, even if the proportion of the spatially structured variance explained by each one of them is quite low.

Figure 2 shows the maps of the indices obtained by classifying the scores of each factor into three isofrequency classes, called low, medium and high. All the maps look quite variable, characterised by many spots of contrasting values, so that it is very difficult to disclose some common patterns of spatial dependence remaining stable over the years.

At a visual inspection, the maps (Figure 2) do not reveal a sensible spatial association from one year to another one, which means that temporal variation, related to meteorological pattern, exceeded spatial variation in the three years. However, to make these comparisons more objective, we calculated two contingency matrices (not shown) to assess the spatial shift of the classes from the 2008 vintage to 2009 vintage and from the 2009 vintage to 2010 vintage. The results show that the classes high, low and medium corresponding to the 2008 vintage remained stable in the corresponding classes (high, low and medium) of the 2009 vintage at the percentages of 37.40, 32.91 and 34.14, respectively, whereas the resting part moved to the other classes. About 34%f the class high was transformed into low class and 31% of low class into high class. As for the transition from the 2009 vintage to 2010 vintage the high class remained stable for

41.39% and moved to low class for 26.52%, whereas the low class remained stable for 46.65% and moved to high for 31.95%. In synthesis, the results show that the overall temporal stability of the classes between the vintages 2008-2009 and 2009-2010 was about 35% and 45%, respectively. These results are confirmed also by the values of k statistics, 0.023 and 0.18 for the two cases, respectively, which are extremely low, even if significantly different from zero. Such a low level of spatial association over time can be attributed mainly to the sensitivity of the vineyard to the contingent conditions.

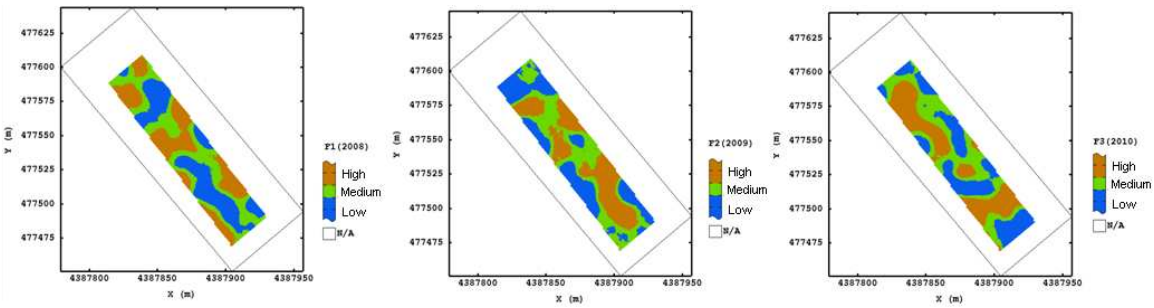


Figure 2: Maps of the first three factors.

4. Concluding remarks

The main objective of this work was to assess the magnitude, structure and persistence in time of the spatial distributions of quantitative and qualitative properties of a vineyard using multivariate geostatistics. In this study we showed that multivariate geostatistics can be used to assess the heterogeneous spatial and temporal distributions within a vineyard and could then be used efficiently in PV. However, to make site-specific management successful, the spatial distribution of vine should be well structured and the temporal persistence high enough. The preliminary results seem to advice against the use of PV in the study vineyard, though the analysis should be repeated over several years in order to reveal valuable recurrent patterns over time.

References

Castrignanò A., Giugliarini L., Risaliti R., Martinelli N., 2000. Study of spatial relationships among some soil physico-chemical properties of a field in central Italy using multivariate geostatistics. *Geoderma*, 97, 39-60.

Bramley R. G. V., and Hamilton, R. P. 2004. Understanding variability in winegrape production system. 1. Within vineyard variation in yield over several vintages. *Australian Journal of Grape and Wine Research*, 10, 32-45.

Bramley, R., Pearse, B. and Chamberlain, P. (2003) Being Profitable Precisely – A case study of Precision Viticulture from Margaret River. *Australian Grapegrower and Winemaker 473a, 84–87.*

Bramley, R. G. V., Proffitt, A. P. B., Hinze, C. J., Pearse, B., & Hamilton, R. P. (2005). Generating benefits from precision viticulture through selective harvesting. In J. V. Stafford (Ed.), Precision agriculture '05 (pp. 891–898). Wageningen, The Netherlands: Wageningen Academic Publishers.

Wackernagel H., 2003. Multivariate Geostatistics: an Introduction with Applications. Springer Verlag Berlin, 3rd ed., 388 pp.

Alternative approaches for probabilistic precipitation forecasting[1]

Francesca Bruno, Daniela Cocchi
Dipartimento di Scienze Statistiche, Università di Bologna, francesca.bruno@unibo.it

Anna Rigazio
Facoltà di Scienze Statistiche, Università di Bologna, annarigazio@gmail.com

Abstract: Bayesian Model Averaging (BMA) and Bayesian Hierarchical Model (BHM) are statistical postprocessing techniques for calibrating precipitation forecast ensembles. BMA is a mixture model of predictive densities, while BHM is a fully Bayesian alternative to BMA. Both techniques are applied on a case-study. BMA is applied to quantitative precipitation, yielding a better calibration than the ensemble in homogeneous areas. For qualitative precipitation, both BMA and BHM forecasts are more calibrated than the ensemble. However, BHM yields a worse performance due to the "shrinkage" effect, that lets the forecasts vary across a small range of values.

Keywords: precipitation forecasting, forecast ensemble, ensemble calibration, Bayesian hierarchical models.

1. Introduction

Short-term weather forecasting is a primary aim in meteorology. Here we consider precipitation, which can be seen either as the binary variable precipitation/no precipitation, or as quantitative precipitation, with reference to precipitation accumulation. The distribution of precipitation accumulation is far from being normal, since it has a positive probability of assuming zero value and is skewed.
Precipitation forecasting has been traditionally regarded as based on deterministic numerical models. A technique for including variability and uncertainty in meteorology is the implementation of ensemble forecasts. An ensemble forecast provides multiple perturbations of numerical predictions differing in the initial conditions and/or the numerical representation of the atmosphere, thereby addressing the two major sources of forecast uncertainty (Raftery *et al*., 2005). If the ensemble forecast consists of multiple perturbations of a single numerical prediction model, its members should be considered undistinguishable and statistically exchangeable, *i.e*., with *prior* equal predictive skills.
Ensemble forecasts are often biased and underdispersive and statistical postprocessing techniques are required to calibrate them. The output is then a predictive Probability Density Function (PDF). In order to warrant a good predictive performance, Gneiting *et al*. (2007) propose to construct probabilistic forecasts according to the diagnostic

[1] Work supported by the project PRIN 2008: New developments in sampling theory and practice, Project number 2008CEFF37, Sector: Economics and Statistics, awarded by the Italian Government.
Thanks are due to Tilmann Gneiting and Thordis Linda Thorarinsdottir for their help and the useful discussions.

approach of sharpness maximization, under calibration. Calibration refers to the statistical consistency between observations and prediction, while sharpness refers to the concentration of the predictive PDF.
Two statistical postprocessing techniques able to calibrate quantitative precipitation ensemble forecasts in presence of exchangeable members are here presented. BMA has been introduced by Raftery *et al.* (2005) and Sloughter *et al.* (2007) as a mixture model of predictive densities, which are themselves mixtures of a discrete component with zero value and a gamma distribution. BMA can be considered as an empirical Bayesian approach, where all parameters are plug-in Maximum-Likelihood estimates. Here we propose a two-level BHM as a fully Bayesian alternative to BMA, extending the model originally introduced by Di Narzo and Cocchi (2010) to precipitation forecasting.

2. Methods

Bayesian Model Averaging (BMA) is a mixture model of the predictive densities of the ensemble members, that accounts for the uncertainty involved in the model selection process. Let y be the future daily precipitation accumulation and let f be an ensemble output consisting of $K>1$ ensemble members forecasts f_k on that day, coming from K different deterministic models. Every ensemble member can be associated with a predictive PDF, $p_k(y|f_k,\theta_k)$, interpretable as the conditional PDF of y, given that f_k is its "best" forecast, while θ_k are the member-specific model parameters. The BMA outcome is a predictive density, that combines predictions under each model in a weighted average:

$$p(y \mid f, x) = \sum_{k=1}^{K} w_k \, p_k(y \mid f_k, \theta_k). \tag{1}$$

For each k, the weight w_k is the model posterior probability over a training period x and reflects the relative predictive skill of each forecast f_k. Under the assumption of exchangeability of the ensemble members, parameter and weights are supposed to be not member-specific; hence we have $\mathrm{w_k} = 1/\mathrm{K}$.
Sloughter *et al.* (2007) propose to model the conditional PDF of precipitation accumulation, $p_k(y|f_k,\theta_k)$, as a finite mixture of a point mass at zero, modeling the probability of non precipitation via logistic regression, and a gamma distribution $g_k(y|f_k,\theta)$, modeling the precipitation accumulation, given that it is greater than zero.
By considering the case of quantitative precipitation and assuming exchangeability, equation (1) can be written as:

$$p(y \mid f, x) = \sum_{k=1}^{K} w_k \left\{ P(y=0 \mid f_k, \theta) I_{[y=0]} + P(y>0 \mid f_k, \theta) g_k(y \mid f_k, \theta) I_{[y>0]} \right\}. \tag{2}$$

BMA is not a fully Bayesian model because the parameter vector θ is not considered as a random variable and, therefore, it is not integrated out. Moreover, the posterior ensemble member probabilities are replaced by Maximum-Likelihood estimates according to an empirical Bayesian approach.
A fully Bayesian version of BMA is offered by a two-level Bayesian Hierarchical Model (BHM) (Di Narzo and Cocchi, 2010). For a BHM, equation (2) is modelled

conditionally on a latent member selection process that every day selects a "best" forecast, as follows:

$$p(y|f,x) = \int_{\theta} \sum_{k=1}^{K} p_k \left(y|f_k, \theta_k, z=k \right) p\left(z=k\right) p\left(\theta|x\right) d\theta \,. \tag{3}$$

Since the ensemble members are assumed exchangeable, is modelled according to i.i.d. discrete uniform distributions and the posterior distribution of the parameters is obtained starting from higher level vague prior distributions.

3. A case-study: Materials and Results

The dataset analyzed contains 24-hour precipitation accumulation data, measured in millimetres. The data consist of observations and their 48-hour ahead forecasts. Both the observed and the forecasted data refer to the period from January 1, 2007, to December 31, 2007. Forecast data are obtained from the deterministic ensemble COSMO Limited-area EPS (COSMO-LEPS), developed by "Azienda Regionale Prevenzione e Ambiente della Regione Emilia Romagna – Servizio Idro-Meteo-Clima" (ARPA-SIM), which includes sixteen exchangeable members. Observed data come from 321 meteorological stations located in the Emilia-Romagna Region. Observations and forecasts concur at each location, after interpolating every meteorological station value to the model grid.

Here both statistical postprocessing techniques are applied on a random sample of 14 grid points. Fig 1. shows the spatial distribution of the mean and the standard deviation of the precipitation accumulation. The two distributions are partitioned according to their quartiles: the figure clearly shows the presence of four spatial clusters . In particular, the stations corresponding to the higher quartiles (red and yellow dots) are placed in the mountains and in the hills, while stations below the median (green and blue dots) are placed in the Po valley and near the seaside.

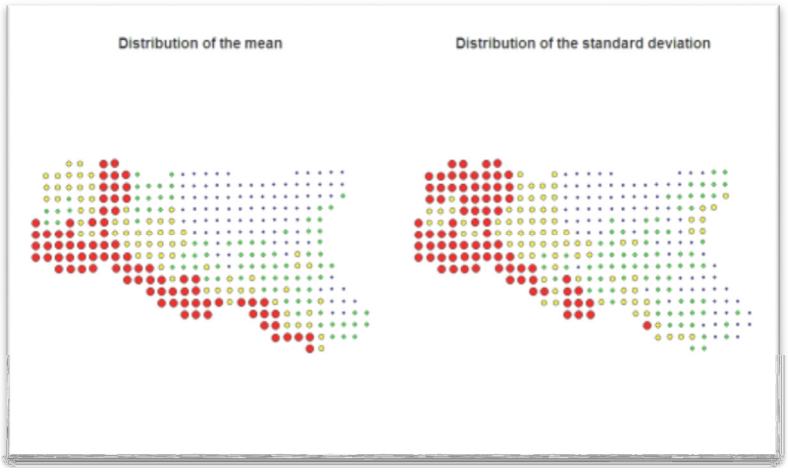


Figure 1: Distribution of the mean and standard deviation of observed precipitation accumulation among the 321 grid points of the dataset.

BMA is useful for global calibration; it is based on the assumption that all meteorological stations are homogeneous, and ignores the possible spatial correlations. This might be too strong an assumption for the Emilia-Romagna Region, which is

characterized by the presence of different climate regimes. This lead us to believe that calibration might be better achieved within subgroups. For this reason a clustering of sites has been adopted following the above criterion. After grouping, stations variability of precipitation accumulation within groups is more homogeneous than before. BMA has been computed separately on random samples selected within the aforementioned clusters. BMA forecasts are far from being calibrated in the two cluster where the variance is larger. The stations of these groups are mostly located in the mountain areas and where the distance from the sea is the farthest, *i.e.* where precipitation is usually less homogeneous. It is known that precipitation is highly affected by local terrain features and some statistical postprocessing techniques for local calibration may yield a better predictive performance in the case of Emilia-Romagna. By considering a random sample of stations only in the clusters where the variability is under the median, BMA yields a better performance than COSMO-LEPS.

BHM is applied on the same sample of stations used for BMA to calibrate qualitative precipitation ensemble forecasts. Both probabilistic forecasts reveal a better predictive performance than COSMO-LEPS forecasts; however, BMA forecasts are more calibrated than BHM ones. Both statistical postprocessing techniques can improve the predictive performance of an ensemble. However, BHM forecasts were more underdisperse than BMA forecasts.

The less satisfactory performance of BMH occurs because this model returns values that are "shrunked" towards the common posterior mean with a very small spread. Since we have chosen vague prior distributions, the posterior results are mostly determined by the observed data, and thus, the posterior common mean of the BHM forecast probabilities is very similar to the mean of the observed frequency of precipitation occurrence. The shrinkage from the maximum likelihood estimates towards the posterior common mean is a typical behaviour in Bayesian hierarchical models. This is usually a valuable characteristics of the Bayesian estimates with respect to the maximum likelihood estimates.

References

Di Narzo A.F., Cocchi D. (2010) A *Bayesian Hierarchical Approach to Ensemble Weather Forecasting*. Journal of the Royal Statistical Society Series C: Applied Statistics, Vol. 59, No. 3, pp. 405-422.

Gneiting T., Balabdaoui F., Raftery A.E. (2007) *Probabilistic Forecasts, Calibration and Sharpness*. Journal of the Royal Statistical, Society Series B: Statistical Methodology, Vol. 69, No. 2, pp. 243-268.

Raftery A.E., Gneiting T., Balabdaoui F., Polakwsky M. (2005) *Using Bayesian Model Averaging to Calibrate Forecast Ensembles*. Monthly Weather Review, Vol. 133, pp. 1155-1174.

Sloughter J.M, Raftery A.E., Gneiting T., Fraley C. (2007) *Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging*. Monthly Weather Review, Vol. 135, pp. 3209-3220.

Comparison of calibration techniques for a limited-area ensemble precipitation forecast using reforecasts

Tommaso Diomede, Chiara Marsigli, Andrea Montani, Tiziana Paccagnella
ARPA-SIMC, HydroMeteorological and Climate Service of the Emilia-Romagna Regional Agency for Environmental Protection, Bologna, Italy, tdiomede@arpa.emr.it

Abstract: The calibration of the precipitation forecasted at high resolution is currently a challenge for the ensemble community working with Limited Area Models. Here, the potential of using reforecasts to achieve this goal was investigated. Different calibration techniques were tested. The impact of the application of these techniques to the precipitation forecasts provided by a Limited-area Ensemble Prediction System was verified over the Emilia-Romagna Region (Northern Italy), Switzerland and Germany. The results revealed a beneficial impact of the calibration process for Switzerland and Germany; rather, no significant improvements were obtained for Emilia-Romagna. As the model error is likely to have a systematic dependence on geography, orography and flow direction, weather-regime dependent correction functions should be generated for improving the calibration strategy.

Keywords: calibration, precipitation forecast, ensemble, reforecasts, COSMO-LEPS

1. Introduction

The calibration of the precipitation forecasted at high resolution is currently a challenge for the ensemble community working with Limited Area Models, especially with respect to the improvement of the forecast skill for rare events. The potential of using reforecasts to achieve this goal has been shown in recent studies (Hamill et al., 2008; Fundel et al., 2010). Reforecasts mean a large dataset of retrospective forecasts obtained by the same model that is run operationally. In the present work, thirty years of reforecast of one member of COSMO-LEPS (the Limited-area Ensemble Prediction System based on the non-hydrostatic limited-area model COSMO) were used for the implementation of the calibration strategy over the Emilia-Romagna Region (Northern Italy), Switzerland and Germany. Three calibration techniques were tested: cumulative distribution function based corrections, linear regression and analogs. The choice of these methodologies is due to the need of improving the quantitative precipitation forecasts (QPFs) provided by COSMO-LEPS, especially as an input to hydrological models. Thus, techniques which enable a calibration of QPFs and not only of the probabilities of exceeding a threshold were selected.

2. Materials and Methods

The calibration strategy was based on the availability of historical forecast and observed rainfall data over the areas under investigation. Thirty years of reforecast of one member of COSMO-LEPS (10 km of horizontal resolution, 40 vertical levels) were run

by MeteoSwiss. One reforecast run with a 90-h lead time was available every three days from 1971 to 2000. This model climatology was used to calibrate forecasts of all lead times, without considering the time dependency of model bias (Fundel et al., 2010). According to the model climatology, the observed precipitation data were collected over the period 1971-2000 for Emilia-Romagna and Switzerland; rather, the observed data over Germany were available only for the period 1989-2000. The rainfall data were interpolated on the model grid points which cover the areas under investigation.
The calibration techniques tested in this work provide corrections based on the Cumulative Distribution Function (hereafter, CDF), the Linear Regression (hereafter, LR) and the Analog method. The described methodologies were used to calibrate each member of COSMO-LEPS. Each calibration function was defined by using the historical data forecasted and observed over each grid point for a specific season.
For the CDF method, the calibrated 24-h QPF was determined by comparing the reforecast and observed CDF curves. The value of the observed data which had the same probability of occurrence of the current 24-h QPF was used as the corrected QPF value. For the LR method, the parameters of the regression line estimated on the basis of reforecast and raingauge historical data were used to correct the current 24-h QPF value. The analog-based methodology was applied using two implementations, which differ from each other for the meteorological field used for the analog search. In the first implementation, the analog search was performed in terms of the similarity of the forecasted precipitation field over the area under investigation. In the second implementation, the analog search was performed in terms of the similarity of the forecasted circulation pattern, evaluated in terms of the geopotential at 700 hPa, 12 UTC (hereafter, Z700), over a spatial domain which is significant for the area under investigation to relate the synoptic circulation to the precipitation at ground. In the following of this paper, the first implementation of the analog-based method is referred to as “ANL” and the second implementation as “anlZ”. For both implementations, for each 24-h lead time, the root-mean-square (rms) differences between each member of the current forecast and each reforecast day were computed (the comparison was carried out among fields coming from the same season). The historical date with the smallest rms difference was chosen as the analog day, then the gridded raingauge recordings of that past day were used as the calibrated QPF.
The impact of the calibration process was verified for 24-h QPFs operationally provided by COSMO-LEPS in the years 2003-2007. The probabilistic verification was carried out in terms of the attributes diagram and the Brier Skill Score (BSS).

3. Results

The results obtained by the application of the calibration strategy are here discussed only for the autumn seasons in the years 2003-2007.
Figure 1 shows the attributes diagram for the lead time day 2. The verification was performed for each model grid point with respect to the ninety-fifth percentile of the climatological distribution of observed 24-h precipitation as threshold for the verified events. For Emilia-Romagna, the raw ensemble has no good reliability, providing overconfident forecasts. Only the calibration based on LR allows an increase of reliability. The weakness of the raw forecast system is more evident over Switzerland (i.e. the raw ensemble lies under the no skill line). The ensembles calibrated by the

CDF, LR and ANL methods show an increase of reliability; nevertheless these ensembles are still overconfident. For Germany, a beneficial impact is provided by the calibration based on LR, whereas a slight increase of reliability results for the ensembles calibrated by CDF and rainfall analogs.
Generally, the calibrated ensembles are still overconfident, especially for high probability values. The calibration based on the analogs of geopotential provides bad performance over all the three study areas. This result reveals that the geopotential at 700 hPa is not a good predictor for the precipitation over the selected areas.

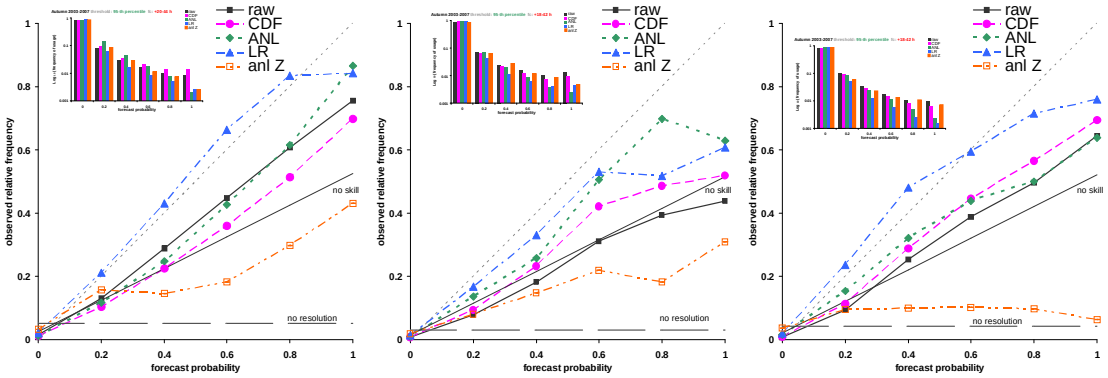


Figure 1: Attributes diagrams for the raw and calibrated ensembles over Emilia-Romagna (left panels), Switzerland (middle panels) and Germany (right panels) in autumn at day 2 lead time, for the 95-th percentile threshold. The inset histograms denote the frequencies of the use of the forecasts for each probability bin.

Figure 2 shows the results obtained in terms of BSS for the autumn season in the period 2003-2007 with respect to the ninety-fifth percentile of the observed climatology as threshold for the verified events. The observed climatology is used as the reference forecast for the computation of the skill score. The calibration process does not provide a beneficial impact on the ensemble QPFs over Emilia-Romagna. Actually, the values of BSS associated to the calibrated ensembles are lower than the BSS of the raw ensemble for all the lead times. The raw ensemble performs worse than climatology over Switzerland, but the calibration process provides the greater amount of skill improvement. With the exception of the anlZ method, the forecasts calibrated by all the methods show a significant increase of BSS values. Even, with respect to climatology, unskillful raw forecasts can be turned into skillful forecasts. In particular, the highest BSS values are provided by the ANL method. For Germany, a beneficial impact is provided by the CDF method for all the lead times; rather, slight improvements are obtained for the ensembles calibrated by LR and rainfall analog only for the longer lead times. Generally, the decay of performance with lead time is evident for the raw and calibrated forecasts.
An additional verification of the calibration process was performed by the coupling of the ensemble precipitation forecasts with an hydrological model. This test was carried out for the Reno river basin, a medium-sized catchment located in the Emilia-Romagna Region. The river hydrograph simulations were carried out for the autumn and spring seasons in the period 2003-2008 by using the distributed rainfall-runoff model TOPKAPI. The results of the coupling were evaluated in terms of missed events and false alarms which would have been issued based on the discharge scenarios driven by the raw and calibrated QPFs, with respect to the exceeding of the warning threshold

defined for the aims of civil protection. The results showed that, on the one hand, a beneficial impact on the reduction of missed events was provided by the calibration performed with the ANL and CDF methods. On the other hand, an increase of false alarms resulted by the application of the two above-mentioned calibration methods, even though this trend is evident for the ANL method only for longer lead times.

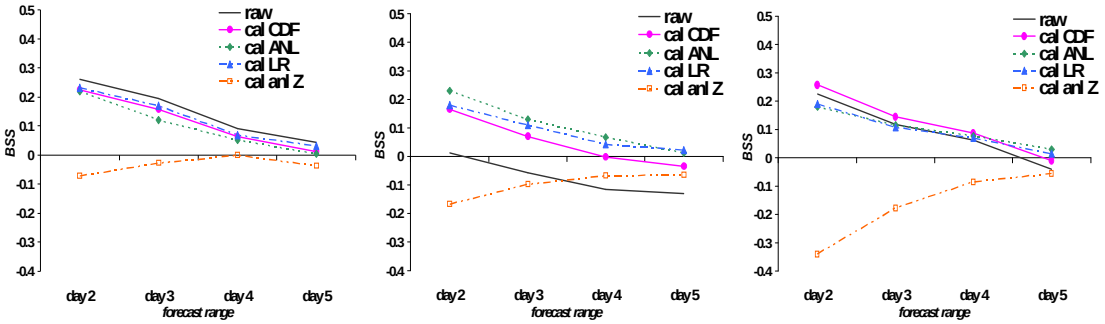


Figure 2: BSS for the raw and calibrated ensembles over Emilia-Romagna (panels on the left), Switzerland (panels in the middle) and Germany (panels on the right) in autumn, as a function of the forecast lead time. Skill at the 95-th percentile threshold.

4. Concluding remarks

The results revealed a beneficial impact of the calibration process over Switzerland and Germany. No significant improvements were obtained over Emilia-Romagna by evaluating the statistical analysis on the calibrated QPFs. The coupling of the QPFs calibrated with the ANL and CDF methods with an hydrological model revealed a beneficial impact of the calibration on the reduction of missed events for a medium-sized catchment (i.e. the Reno river basin) used as a test-bed. The lack of a remarkable improvement, especially over Emilia-Romagna, resulting from the application of the proposed calibration methods suggests the need of defining specific correction functions which should be able to link the model errors to the meteorological situation. Actually, the search for a unique relationship between forecast and observed data hampers to highlight the model errors which are known to have a systematic dependence on geography, orography and flow direction. Therefore, the calibration strategy should be improved by dividing the training sample size in order to pool data which have similar model errors with respect to a given meteorological situation.

References

Fundel F., Walser A., Liniger M.A., Frei C., Appenzeller C. (2010) Calibrated Precipitation Forecasts for a Limited Area Ensemble Forecast System Using Reforecasts, *Mon. Weather Rev.*, 138, 176-189.

Hamill T.M., Hagedorn R., Whitaker J.S. (2008) Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation, *Mon. Wea. Rev.*, 136, 2620–2632.

Functional boxplots for summarizing and detecting changes in environmental data coming from sensors

Elvira Romano, Antonio Balzanella
Dipartimento di Studi Europei e Mediterranei, Seconda Universitá degli Studi di Napoli, elvira.romano@unina2.it

Lidia Rivoli
Dipartimento di Matematica e Statistica, Universitá di Napoli Federico II

Abstract: Nowadays, environmental sensor networks produce a large amount of streaming time series whose storage, manipulation and indexing is impractical. In this work, we propose a new strategy for summarizing and describing this kind of data based on functional data representation. It discovers trends and potential anomalies by using an informative exploratory tool: the functional boxplot. Functional boxplots are introduced for conveying location and variability information. In addition, for detecting and illustrating variation a distance among functional boxplots is used.

Keywords: streaming time series, functional data, functional boxplot

1 Introduction

In a wide range of environmental applications, networks of sensors allow to record huge amounts of temporally ordered data. Often, the sampling frequency is very high and the monitored phenomenon is highly evolving. This involves that traditional temporal data mining methods, based on computationally intensive algorithms and requiring the storing of the whole dataset, become ineffective. Especially there is a remarkable delay between the recording of the data and the analysis results which can impact on decisional processes.

In order to deal with this issue, it is necessary to move from the traditional temporal data mining to the data mining of streaming time series which focuses on processing the incoming data on-line without requiring their storage.

Usually, algorithms for data streams mining update, in incremental and on-line way, the knowledge about data by means of synopses. These provide suitable summaries which are substantially smaller than their base dataset and allow to discard the data once they have been processed. In literature, several summarization techniques for streaming time series have been proposed (a wide review is available in (Mitsa T., 2010)). Some of these transform a streaming time series into a new one of reduced dimensionality, others use sampling, sketches, histograms.

In this paper we introduce an intuitive tool for visualizing and summarizing

the behavior of multiple streaming time series, the Functional Box Plot (FBP). Originally defined for functional data, it is considered as a variable and used as synthesis of batches of the incoming multiple streaming time series. The monitoring of the evolution of the data has been performed through the comparison of the FBP variables using an appropriate distance measure, rather than analyzing the incoming recordings.

2 The three steps strategy

Let $y_i(t),\ \ i = 1, \ldots, n, t \in [1, \infty]$ a set of streaming time series made by real valued ordered observations of a variable $Y(t)$ in n sites, on a discrete time grid.

Our aim is to summarize and describe their changes in a streaming fashion by means of a comparison of functional boxplot variables. Functional boxplots are an informative explorative tool for functional data. We use them as variables of synthesis for the set of n streaming time series splitted in non overlapping windows and opportunely approximated by functional data. With this scope a three steps strategy is proposed.

The first step consists in splitting the incoming parallel streaming time series into a set of non overlapping windows $W_j, j = 1, \ldots, \infty$, that are compact subsets of T having size $w \in \Re$ and such that $W_j \bigcap W_{j+1} = \emptyset$. The defined windows frame for each $y_i(t)$ a subset $y_i^{w_j}(t)\ \ t \in W_j$ of ordered values of $y_i(t)$, called subsequence.

Following the FDA approach, we consider each subsequence ${y_i}^{w_j}(t)$ of $y_i(t)$ the raw data which includes noise information (Ramsay, J.E., Silverman, B.W., 2005). Then we determinate a true functional form $f_i^{w_j}(t)$, we call functional subsequence, which describes the trend of the flowing data, by using smoothing spline functions. For each W_j we have that all the subsequences $y_i^{w_j}(t)\ \ i = 1, \ldots, n$ follow the model:

$$y_i^{w_j}(t) = f_i^{w_j}(t) + \epsilon_i^{w_j}(t),\ t \in W_j\ \ i = 1, \ldots, n \tag{1}$$

where $\epsilon_i^{w_j}(t)$ are residuals with independent zero mean and $f_i^{w_j}(\cdot)$ is the mean function which summarizes the main structure of $y_i^{w_j}(t)$.

In a second step since we need to have a summary of the batched streaming time series, we compute functional boxplot variables for each batch. Functional boxplot(box-and-whisker diagram or plot) is an informative graphically tool for depicting functional data through their five-functions summaries. We consider them as a kind of quantitative variables in the functional setting.

In functional data analysis two different definition of boxplot exist. A first one makes use of the first two robust principal component scores, Tukey data depth and highest density regions (Hyndman R.J., Shang, H.L., 2010); a second one is based on center outward ordering induced by band depth for functional data (Sun Y., Genton G., 2011). We makes use of the second boxplot definition, that is a natural extension to the classical boxplot. It is defined starting by a concept which allows to order curves from center outward: the band depth BD (Lopéz-Pintado and Romo 2009).

Let $f_i^{w_j}(t), i = 1, \dots n$ be the collection of functional subsequences in a window W_j, $G(f_i^{w_j}) = \left\{(t, f_i^{w_j}(t)) : t \in W_j\right\}$ be the graph of the function $f_i^{w_j}(t)$, and

$$B\left(f_{i_1}^{w_j}, \dots, f_{i_k}^{w_j}\right) = \left\{\left(t, g_i^{w_j}(t)\right) | t \in W_j, \min_{r=1,\dots,k} f_{ir}^{w_j}(t) \leq g_i^{w_j}(t) \leq \max_{r=1,\dots,k} f_{ir}^{w_j}(t)\right\} \quad (2)$$

be the band in R^2 delimited by the k different curves $\left(f_{i_1}^{w_j}, f_{i_2}^{w_j}, \dots, f_{i_k}^{w_j}\right)$, obtained by computing the minimum and the maximum values for all t. Let $BD_n^{(m)}$ be the portion of bands obtained by $m = 1, \dots, M$ different curves containing the whole graph of $f_i^{w_j}(t)$ expressed by

$$BD_n^{(m)}(f_i^{w_j}) = \binom{n}{m}^{-1} \sum_{1 \leq i_1 \leq i_2 \leq \dots \leq i_m \leq n} I\left\{G\left(f_i^{w_j}\right) \subset B\left(f_{i_1}^{w_j}, f_{i_2}^{w_j}, \dots, f_{i_m}^{w_j}\right)\right\} \quad m \geq 2 \quad (3)$$

where $I\{\cdot\}$ denote the indicator function.

Thus the band depth $BD_{n,M}(f_i^{w_j}(t))$of any of these function $f_i^{w_j}(t)$ is defined as

$$BD_{n,M}(f_i^{w_j}) = \sum_{m=2}^{M} BD_n^{(m)}\left(f_i^{w_j}(t)\right) \quad M \geq 2 \quad (4)$$

Especially let $f_{[i]}^{w_j}(t)$ denote the sample of functional subsequence associated to the ith largest band depth value, the set $f_{[1]}^{w_j}(t) \dots, f_{[n]}^{w_j}(t)$ are order statistics, with $f_{[1]}^{w_j}(t)$ the median curve, that is the most central curve (the deepest), and $f_{[n]}^{w_j}(t)$ is the most outlying curve. Moreover the central region of the boxplot is defined as

$$C_{0.5} = \left\{(t, f^{w_j}(t)) : \min_{r=1,\dots,[n/2]} f_{[r]}^{w_j}(t) \leq f^{w_j}(t) \leq \max_{r=1,\dots,[n/2]} f_{[r]}^{w_j}(t)\right\} \quad (5)$$

where $[n/2]$ is the small integer not less than $n/2$. The border of the 50% central region is defined as the envelope representing the box of the classical boxplot.

Based on the center outwards ordering induced by band depth for functional data, the descriptive statistics of such functional boxplots FBP are: the upper $f_{[u]}^{w_j}(t)$ and lower $f_{[l]}^{w_j}(t)$ curves (boundaries) of the central region, the median curve $f_{[1]}^{w_j}(t)$ and the non-outlying minimum $f_{[b_{min}]}^{w_j}(t)$ and maximum boundaries $f_{[b_{max}]}^{w_j}(t)$.

For each window we have a FBP variable that is considered as a variable compound of five sub functions with the following structure:

$$\left\{f_{[u]}^{w_j}(t), f_{[l]}^{w_j}(t), f_{[1]}^{w_j}(t), f_{[b_{min}]}^{w_j}(t), f_{[b_{max}]}^{w_j}(t)\right\} \quad (6)$$

The third and latest step, consists in monitoring the evolution of the multiple data streams by comparing functional boxplot variables. With this aim we introduce a distance measure between a pair of FBP variables. It is a Manhattan distance which extends the distance for classical boxplot introduced in Arroio J., Mat C., Roque A. (2006) to functional boxplot variables. It is computed by considering that

each couple of correspondent functions is compared on the same time interval W by means of a transformation of the functions domain. Thus, the Manhattan distance between a pair of functional boxplot FBP_1, FBP_2 opportunely shifted is:

$$\begin{aligned} d(FBP_1, FBP_2) &= \left| \int_{t \in W} (f'^{w_1}_{[u]}(t) - f'^{w_2}_{[u]}(t))dt \right| + \left| \int_{t \in W} (f'^{w_1}_{[l]}(t) - f'^{w_2}_{[l]}(t))dt \right| + \\ &+ \left| \int_{t \in W} (f'^{w_1}_{[1]}(t) - f'^{w_2}_{[1]}(t)dt) \right| + \left| \int_{t \in W} (f'^{w_1}_{[b_{min}]}(t) - f'^{w_2}_{[b_{min}]}(t))dt \right| + \\ &+ \left| \int_{t \in W} (f'^{w_1}_{[b_{max}]}(t) - f'^{w_2}_{[b_{max}]}(t))dt \right| \end{aligned}$$

where $f'^{w_j}_{[u]}(t), f'^{w_j}_{[l]}(t), f'^{w_j}_{[1]}(t), f'^{w_j}_{[b_{min}]}(t), f'^{w_j}_{[b_{max}]}(t)$ are the descriptive functions of the shifted FBP. The synthesis obtained by the FBP allows to have a description of batched streaming time series that can be compared on different time interval, thus this distance can be applied also on different and non consecutive time windows.

3 Concluding remarks

In this paper we have introduced a new strategy for summarizing multiple streaming time series and for monitoring their evolution. Unlike approaches existent in streaming time series literature, we have introduced a tool able also to provide an intuitive graphic summarization of data.

We have performed several tests on climate data in order to assess the effectiveness of the method. Preliminary results are encouraging.

References

Arroyo J., Mat C., Roque A. (2006) *Hierarchical clustering for boxplot variables*, Studies in Classification, Data Analysis, and Knowledge Organization, Part II, 59-66.

Hyndman R.J., Shang, H.L. (2010) Rainbow plots, bagplots and boxplots for functional data, *Journal of Computational and Graphical Statistics*, 19(1), 29-45.

Lopez-Pintado S., Romo, J. (2009). On the Concept of Depth for Functional Data. Journal of the American Statistical Association, 104, 718-734.

Mitsa T. (2010) Temporal Data Mining. Data Mining and Knowledge Discovery Series. Chapman & Hall/CRC

Ramsay, J.E., Silverman, B.W. (2005) *Functional Data Analysis* (Second ed.).Springer.

Sun Y., Genton M.G. (2011) Functional boxplots. *Journal of Computational and Graphical Statistics*. To appear.

Information, advice, friendship, notes and trust network: evidence on learning from classmate [1]

Emma Zavarrone

Iulm University, Milan, Italy, emma.zavarrone.iulm.it

Agnese Vitali

Bocconi University and Dondena Centre, Milan, Italy; agnese.vitali@unibocconi.it

Abstract: This paper contribute to the literature on the influence of network structure and performance of university students over time. We move from the assumption that students' school performance is influenced by: friendship, exchange of general information about the course, contents, lecture notes and trust networks. Social influence has been modeled through SARAR model and several spatial weight matrixes W and M have been compared.

Keywords: SARAR model, social influence, social network analysis

1. Introduction

A number of studies in social network analysis, economic and sociology have recently focused on the association between friend networks (or peer effect) and school performance.

In this paper we aim at contributing to the literature on the influence of network structure and performance of university students. We hypothesize that class mates can develop four different types of relationships. Existing literature usually focuses on two main types of such relationships, namely friendship and study networks. In this paper we explore also the role of two other relationships which might be associated with students' performance: the exchange of general information and the exchange of lecture notes. We measure school performance by the means of students' University Human Capital (UHC).

We test whether the ego's UHC is influenced by the UHC of the subgroup of class mates with which he/she has a relationship of one of the four kinds and which of these relationships has the higher marginal effect on UHC (i.e., we investigate whether UHC is influenced by the UHC of study-network members, as well as by the UHC of friendship-, information- and lecture notes-network members) (*hypothesis 1*). Also, we inspect if students' UHC is influenced by unobserved characteristics common to the ego's networks' structure (*hypothesis 2*). We will test the hypothesis that high-performance students tend to relate themselves with other high-performance students, to isolate low-performance students and to have less free time to spend hanging out with friends (*hypothesis* 3).

2. Materials and Methods

We developed an ad-hoc survey in which students in a given class are asked to detail the structure of four different networks to which they belong. Respondents are master students in the age range 22-23, attending the Statistics course during their first year of a two-year master degree at Iulm University in Milan, Italy.

Students are asked whether, in order to prepare the Statistics exam, they studied on their own. If they did not, they are asked to identify the class mates with whom they studied. We consider these peers as members of the ego's study network. Students are also asked whether they have class mates with which they get together outside the university environment, and if they do, we ask to identify them and we consider them as members of the ego's friendship network. In order to identify the information network and the lecture notes network, we look at class mates whom the ego considers a reliable source of information for what concerns the Stats course and with which he/she exchange/compare his/her notes, respectively. The structure of all the four networks we consider are such that relationships do not necessarily need to be reciprocal. Students are surveyed twice: before the mid-term exam and before the final exam at the end of the course. Student UHC is then measured using the difference between the student's grade obtained at time 1 and the grade he/she obtained at time 0. Both grades are expressed in thirtieths (minimum for sufficiency is 18), a UHC equal to 0 is interpreted as no change in performance between time 1 and time 0, while a positive (negative) UHC is interpreted as increased (decreased) performance between time 1 and time 0.

In addition to information relating to the structure of the four networks discussed above, the survey also collects information on the students' field of education during their bachelor studies, whether their university career took place in the same University in which they are surveyed, and if this is not the case, in which university they took their bachelor. Further, the students are asked whether, during their university studies, they took a Stats class, and if this was the case, they are asked to specify which class it was. Finally, a question is asked to identify who the ego subjectively perceives as the central subject among his/her class mates ("You are the person located in the bottom part of this picture. Could you specify, among your class mates, the initials of the person in the upper left of the picture?"). This question does not refer to one particular network, rather it aims at catching the ego's perception about the central subject among his school mates, in general.

In order to test our three assumptions, we employ respectively the models:

1) the spatial lag model, $UHC = \rho_1 WUHC + X'\beta + \varepsilon;\ \varepsilon \sim N(0,\sigma_\varepsilon^2 I)$

2) the spatial error model, $UHC = \mathbf{X}\beta + \varepsilon;\ \varepsilon = \rho \boldsymbol{W} \varepsilon + \nu\ ;\ \nu \sim N(0,\sigma_\nu^2 I)$

3) the spatial auto-regressive auto-regressive model (SARAR):

$UHC = \rho_1 WUHC + X'\beta + u;\ u = \rho_2 Mu + \varepsilon$

The independent variables (gender and mark of previous statistics ability) are the same for all models; the coefficient ρ_1 measures the spatial autocorrelation in the dependent variable i.e. *a spatial lag* (Cliff et al., 1973; Leenders, 2002); if this coefficient is positively significant, there is evidence of spatial autocorrelation in UHC or, in other words, that students belonging to the same network tend to have similar grade differentials over the two time periods. The coefficient ρ_2 measures instead the spatial autocorrelation in the error term; if this coefficient is positively significant we interpret that there are common unobserved factors influencing all members of the same network (i.e., unobservable factors will have an effect on the network member's UHC to which they are related, but also on the UHC of his/her peers). The spatial weight matrixes W and M, which need not be equal, are non-stochastic spatial weight matrixes which take into account the neighbouring structure of the students, such that their entries are non-null (i.e., two students are neighbours) if the students belong to the same network.

For each weight matrix, we also define a different set of weights in order to assess the robustness of the results found, on the basis of different weight structures. To this aim, in the first place weights will be defined in such a way to assign the same weight to all members of a given network (thus weights will be proportional to the number of people belonging to the specified network). In the second place, weights will be defined to assign more weight to the peer who is central in the network.

Other model assumptions require that the spatial autoregressive ρ_1 and ρ_2 coefficients are bounded in absolute value (i.e. $|\rho_1|<1$ and $|\rho_2|<1$), ε_i is independently and identically normally distributed with zero mean and variance to be estimates. The model can be estimated via Maximum Likelihood or following a GMM procedure. Due to the narrowness of our sample size (n=41), we rely on the second approach. We test the significance of the two spatial autocorrelation coefficients using Lagrange Multiplier Tests.

3. Results

The final sample is constituted of 41 students Table 1 shows some socio-demographic information of the our sample: it is not surprising that male students represent only a minority (34%) at Iulm university (it is well known that a gender difference exists when the field of study is concerned; in particular, women are more often found in humanistic subjects).

Variable	Mean/Pro	Std. Dev.
Final grade in Stats	22	5.2
Sex (prop. of men)	34%	-
BSc in different	48%	-
Ever studied Stats	56%	-

Table 1: Descriptive statistics

For the lag spatial model we find a network effect on performance among class mates. The trust network exerts the most powerful effect on students' performance, followed by the friend and study networks. The Figure 1 summarize these results by the mean of a graphical representation. On the y-axis there is the magnitude of the spatial coefficient. Then we group models by weight matrix, so in the first case, using the friend network, the spatial weight matrix can be defined on the basis of 7 different criteria. And the same goes for each of the other 4 types of networks.
For the other two models we don't find a statistically significant effects and we are testing misspecification procedure on ρ_1 *and* ρ_2.

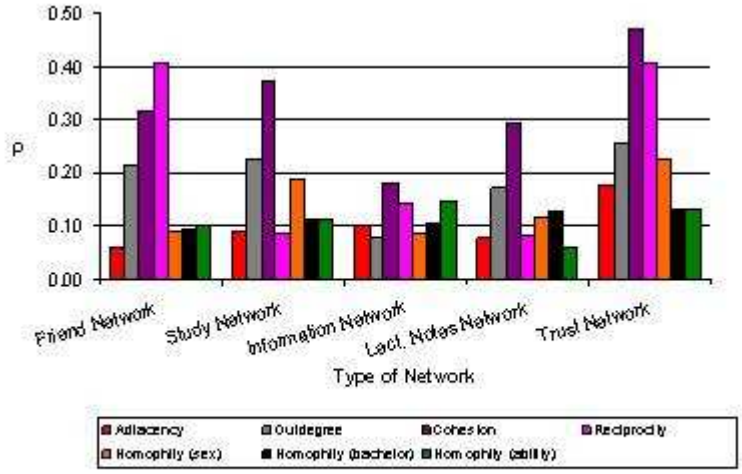


Figure 1: Comparison of ρ coefficient (*p-value* ≤ 0.05) in five networks

4. Concluding remarks

Using a set of different, equally theoretically-grounded weight matrices we show that:

i. in some cases results are robust to different specifications of *W*,
ii. however, in some other cases parameter estimates –hence conclusions– based on autocorrelation models can change according to the chosen specification of *W*,
iii. the network structure need to be translated into a meaningful and theory-guided choice of weight matrix (Leenders, 2002).

References

Anselin, L. (1988) *Spatial Econometrics: Methods and Models*. Boston, Kluwer Academic.

Cliff, A.D. and Ord, J.K. (1973 *Spatial Autocorrelation*. London: Pion Limited.

Leenders, R. Th. A. J. (2002) Modeling social influence through network autocorrelation: Constructing the weight matrix. *Social Networks*, 24, 21-47.

Optimal spatial design for air quality measurement surveys: what criteria ?

Thomas Romary[1], Chantal de Fouquet[1]
1 Geostatistics team, Geosciences centre, Mines ParisTech, 35 rue Saint Honoré, 77305 Fontainebleau, France, thomas.romary@mines-paristech.fr

Laure Malherbe[2]
2 Institut National de l'Environnement Industriel et des Risques (INERIS), Direction des risques chroniques, Parc Technologique Alata, 60550 Verneuil-en-Halatte, France

Abstract: In this work, we present a spatial statistical methodology to design benzene air concentration measurement surveys at the urban scale. In a first step, we define an a priori modeling based on an analysis of data coming from previous campaigns on two different agglomerations. More precisely, we retain a modeling with an external drift which consists of a drift plus a spatially correlated residual. The statistical analysis performed leads us to choose the most relevant auxiliary variables and to determine an a priori variogram model for the residual. An a priori distribution is also defined for the variogram parameters, whose values appear to vary from a campaign to another. In a second step, we optimize the positioning of the measuring devices on a third agglomeration according to a Bayesian criterion. Practically, we aim at finding the design that minimizes the mean over the urban domain of the universal kriging variance, whose parameters are based on the a priori modeling, while accounting for the prior distribution over the variogram parameters. Two optimization algorithms are then compared: simulated annealing and a particle filter based algorithm.

Keywords: Optimal Design, Geostatistics, External Drift Kriging

1 Introduction

Mapping air pollution as precisely as possible is a major issue for French Local Air Quality Monitoring Agencies (the AASQAs) both for regulatory and information purposes and for public health concerns. Seasonal or annual average concentration maps can be obtained from passive sampling data collected at a large number of sites across the area of interest. The AASQAs regularly carry out such sampling surveys over various areas at various scales. Given those considerations, they have to design sampling schemes so that resulting concentration maps will fulfill precision criteria.
The interpolation is performed by kriging, see *e.g.* [1], and [2] for an application in atmospheric sciences. With its internal quantification of spatial variability through the covariance function (or variogram), kriging methodology can produce maps of optimal

predictions and associated prediction error variance from incomplete and possibly noisy spatial data. Kriging also provides a prediction error variance that can be seen as a criterion quality of the resulting maps. As it only depends on the spatial repartition of the points over the domain, it is a straightforward criterion for the quality of a sampling design, once a geostatistical model has been fitted on the phenomenon under study. This work completes and extend [3].

2 Materials and Methods

The proposed methodology consists of three steps: an estimation step, the definition of a quality criterion for the sampling design and an optimization step.

2.1 Estimation

A geostatistical analysis of the data collected during previous surveys is performed in order to set up the model (covariates and covariance) that will be used when applying the optimization method to another area. Data from benzene sampling surveys conducted in two French cities (Lille and Reims) have been used to fit the geostatistical model, which is made of a drift plus a spatially correlated residual:

$$Z(x) = \beta_0 + Y'(x)\beta + S(x), \tag{1}$$

where Z is the benzene concentration variable, $x \in \mathcal{X} \subset \mathbb{R}^2$ is the spatial coordinate, Y is the matrix of auxiliary variables exhaustively known on $\mathcal{X}$, $'$ is the transpose operator, β is a vector of parameters and $S(x)$ is a centered, spatially correlated residual.

2.2 Criterion building

The criterion to optimize is defined from the set up model by the integral over the domain under study of the weighted prediction error variance:

$$O(\eta) = \frac{1}{|\mathcal{X}|} \int_{\mathcal{X}} \mathbb{V}(Z(x) - \widehat{Z}(x)) w(x) dx, \tag{2}$$

where $|\mathcal{X}|$ is the area of $\mathcal{X}$. Practically, this integral is evaluated on a grid discretizing the domain $\mathcal{X}$. A non uniform weight function $w(x)$ can be designed to obtain a more accurate mapping in some areas, for instance in function of auxiliary variables values. When some parameters of the model cannot be fitted accurately, we can associate them an a priori distribution, then a Bayesian version of (2) can be considered:

$$O_{Bayes}(\eta) = \frac{1}{|\mathcal{X}|} \int_{\times} \int_{\mathcal{X}} \mathbb{V}(Z(x) - \widehat{Z}(x)|\theta) w(x) p(\theta) dx d\theta, \tag{3}$$

where $\theta \in \Theta$ is the set of uncertain parameters and $p(\theta)$ is its a priori distribution.

2.3 Optimization

Once the model to use has been clearly identified, we have to optimize (3) on a discretization of $\mathcal{X}$. Optimizing (3) on a large grid is a hard combinatorial problem. Therefore, we rely on heuristics algorithm to perform the global optimization: a simulated annealing scheme and a interacting particle algorithm.

3 Results

A third French agglomeration (Bordeaux) is taken as application case. The performances of both algorithms of are compared, in terms of optimization quality and computing time. We also show how the method can be used to dimension the network of passive samplers.

4 Concluding remarks

The current work has been carried out with the aim of supplying scientific and technical support to the French local air quality monitoring agencies. For the moment it has been applied to benzene sampling over urban areas but it can be extended to other pollutants such as NO2 and to larger spatial domains like regions.

References

[1] Chilès, J. P., and Delfiner, P. *Geostatistics, Modeling Spatial Uncertainty.* John Wiley & Sons, New-York, 1999.

[2] de Fouquet, C., Gallois, D., and Perron, G. Geostatistical characterization of the nitrogen dioxyde concentration in an urban area. part 1: spatial variability and cartography of the annual concentration. *Atmospheric Environment 41* (2007), 6701–6714.

[3] Romary, T., de Fouquet, C., and Malherbe, L. Sampling design for air quality measurement surveys: An optimization approach. *Atmospheric Environment In Press, Corrected Proof* (2011), –.

Point–process statistical analysis for the ECMWF Ensemble Prediction System

Fabrizio Nerozzi
ARPA Emilia–Romagna, Servizio IdroMeteoClima, fnerozzi@arpa.emr.it

Abstract: The possibility of applying mathematical tools of point–process statistics to the ECMWF Ensemble Prediction System (EPS) is exploited in this work, in order to provide a different way to reduce ensemble information. The first two empirical orthogonal functions enable to represent 5–day ensemble forecasts as point processes in a plane. These planar representations are hence compared to a sample of Gaussian random point patterns, obtained by a Montecarlo method. The estimations of the nearest–neighbour distribution function and of the reduced second order momentum function for point processes relative to the ensemble predictions are in good agreement with the corresponding estimations of Gaussian random point processes.

Keywords: Ensemble predictions, Principal Component Analysis, Point–process statistics

1 Introduction

Ensemble predictions appear to be the only feasible method to predict the evolution of the atmospheric probability distribution function beyond the range in which error growth can be prescribed by linearized dynamics (Molteni *et al.*, 1996).

However, the large amount of information contained in the ECMWF Ensemble Prediction System (EPS) can be hardly managed in the whole, and two different strategies, clustering and tubing, are adopted for reducing the 51 EPS members to few alternative scenarios. For clustering, "similar" EPS forecasts are collected in clusters, whose probabilities of occurrence are provided by the cluster sizes (Molteni *et al.*, 2001). As concerns the tubing technique, this consists in an averaging of all ensemble members close to the ensemble mean, while the excluded members are grouped together in a number of tubes. Each tube is represented by its most extreme member belonging to it.

One of the principal shortcomings of the clustering technique is the empirical distribution of the ensemble members. Although tubing allows a better visualization of the most different scenarios in the ensemble than clustering, tubes do not provide probabilities of occurrence. In order to provide a different way to condense information, which tries to overcome these shortcomings of clustering and tubing techniques, it is here exploited the possibility to represent ensemble forecasts as a finite set of random points distributed in a plane. In particular, it is tested the

hypothesis according to which these ensemble point processes can be treated statistically equivalent to Gaussian random point processes.

2 Materials and Methods

2.1 The Principal Component Analysis

The Principal Component Analysis (PCA) enables to transform a data set, characterized by a large number of variables in a new one, where the number of variables is highly reduced. The new variables are calculated as the eigenvectors of the covariance matrix and they are orthogonal among them (Preisendorfer, 1988).

For each day of the whole meteorological winter season 2006–2007, starting from the 1st December 2006 to the 28th February 2007, it has been computed the covariance matrix of the 51 ECMWF EPS 5–day forecasts at 12 UTC of the 500 hPa geopotential height. The geopotential height is a meteorological field, here defined over a regular grid at 1 degree of resolution and covering the European area (33N–74N; 27W–45E). Then, for each one of the 51 ensemble members the PCA technique has been applied, and the first two normalized principal components have been considered. The explained variance by these first two principal components ranges from 46% to 69% of the total variance.

Eventually, in order to represent the ensemble forecasts as a a random point–process (hereafter called EPS point–process), for each one of the 90 winter days the 51 ensemble members are represented as single points lying over the plane formed by the first two PCA eigenvectors, whose coordinates are provided by the first two principal components of the ECMWF EPS members.

2.2 The point–process statistics

The Gaussian random point–process is here taken as reference model. In particular, for each winter day 199 bidimensional Gaussian random point–processes with 51 points, zero mean and variance equal to 1, have been simulated by a Montecarlo method. Hence, for the corresponding EPS point–process and for these 199 simulations the nearest–neighbour distribution function $D(r)$ has been defined computing the distance from the analysis point (the "observed" 500 hPa geopotential height reduced to the first two principal components), chosen as the arbitrary event, to its nearest event belonging to each one of the 200 random patterns. Analogously, the derivative L of the reduced second order momentum, K function, is computed counting for each point pattern the number of events within a distance r from the analysis point.

The nearest–neighbour distribution function $D(r)$ describes the probability that distance from a randomly chosen event to its nearest event is less than or equal to $r > 0$. This function can be heuristically estimated from the observed pattern:

$$\hat{D}(r) = \frac{\sum_{i=1}^{n} I(r_{i,A} \leq r, d_i > r)}{\sum_{i=1}^{n} I(d_i > r)} \tag{1}$$

where d_i denotes the distance of the event from the nearest boundary of the closed set A and $r_{i,A}$ is the distance from the nearest event in A (Cressie, 1991).

The K function uses information in the pattern over a wide range of scales than the nearest–neighbour distribution function. Its definition is related to the number of extra events within distance r from an arbitrary event. Estimating of K from an observed pattern in a bounded $A \subset \Re^2$ is complicated by edge effects. Here the Ripley's edge–corrected estimator is considered (Cressie, 1991):

$$\hat{K}(r) = \frac{1}{n\lambda} \sum_{i=1}^{n} \sum_{j=1'}^{n} w(\mathbf{s}_i, \mathbf{s}_j)^{-1} I(\|\mathbf{s}_i - \mathbf{s}_j\| \leq r) \tag{2}$$

The estimator $\hat{K}$ is approximately unbiased provided that the n events are approximately independent. Estimates of the derivative of the K function, $\hat{L}(r)$, are computed by the formula (Stoyan *et al.*, 1987):

$$\hat{L}(r) = \sqrt{\frac{\hat{K}(r)}{\pi}} \tag{3}$$

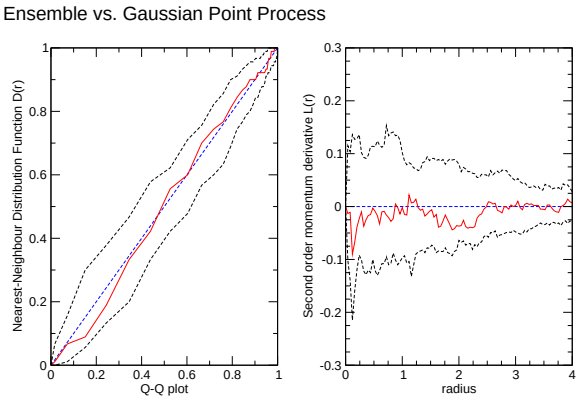


Figure 1: EPS against Gaussian Point–Processes: nearest–neighbour distribution function Q–Q plot (left panel), and the derivative of the K function (right panel).

3 Results and Concluding remarks

The possibility of applying mathematical tools of point–process statistics to the ECMWF Ensemble Prediction System has been exploited in this work, in order to

look for a different way to condense the large amount of ensemble information. It has been proved the first two empirical orthogonal functions enable to represent about, or more, 50% of the ensemble spread. Therefore the ensemble forecasts have been represented as point processes in the plane of the first two PCA eigenvectors and compared to Gaussian random point processes, obtained by a Montecarlo method and considered as reference models.

In figure 1, it is reported on the left side the Q-Q plot, where the quantiles of the median of the 199 nearest–neighbour functions relative to 90 Gaussian random point processes are in the abscise axis. In the ordinate axis there are the quantiles of the nearest–neighbour function relative to the 90 ensemble point processes (continuous red line), the median (dashed blue line), the minimum and maximum (dashed black lines) of the 199 nearest–neighbour functions. On the right side, it is instead reported the derivative of the K function relative to the 90 ensemble point processes minus the median of the 199 simulations (continuous red lines). Analogously, the confidence interval is represented by the minimum and maximum of the 199 simulations, again subtracted by the median (dashed black lines).

The good agreement between ensemble and Gaussian random point processes, in terms of the nearest–neighbour distribution function and of the reduced second order momentum function estimations, coming out of the present work, could render plausible to consider the probability distribution function of ensemble members as asymptotically normal.

References

Cressie N. (1991) *Statistics for spatial data*, J. Wiley & Sons, New York.

Molteni F, Buizza R., Palmer T. N., Petroliagis T. (1996) The ECMWF Ensemble Prediction System: Methodology and validation. *Q. J. R. Meteorol. Soc. 122, 73–119.*

Molteni F., Buizza R. (1999) Validation of ECMWF Ensemble Prediction System using empirical orthogonal function. *Mon. Wea. Rev., 127, 2346–2358.*

Molteni F., Buizza R., Marsigli C., Montani A., Nerozzi F. and Paccagnella T. (2001) A strategy for high–resolution ensemble prediction. Part I: Definition of Representative Members and Global Model Experiments. *Q. J. R. Meteorol. Soc., 127, 2069–2094.*

Preisendorfer R. W. (1988) *Principal component analysis in meteorology and oceanograpy*, Curtis D. Mobley, New York.

Stoyan D., Kendall W. S., Mecke J. (1987) *Stochastic geometry and its applications*, J. Wiley & Sons, New York.

Combining geostatistics and process-based water quality model to improve estimation along a stream network. Example on a stretch of the Seine River

de Fouquet C., Polus-Lefèbvre E., Flipo N., Poulin M.
Centre de Géosciences / Equipe Géostatistique, Ecole des Mines de Paris, France; e-mail: chantal.de_fouquet@ensmp.fr

Abstract: Models that estimate pollutant concentrations in streams can roughly be classified into two categories. Physically-based models tend to reproduce processes and provide dense information, but mostly does not suit the measurements. Stochastic models are based on observations, but the monitoring network usually provides too few measurements for a relevant estimation. This paper aims at combining both approaches to improve water quality characterization.

First a comparison of measurements and model outputs is performed. Then a geostatistical multivariate estimation method is used to combine them and provide a measurement interpolation based on process-based model outputs, with joint uncertainty quantification. The reasoning is applied to nitrate and dissolved oxygen concentrations.

Landscape impacts of photovoltaic plants on the ground: a case-study through the application of rendering techniques

N. Robles*, R. Primerano, V. Perrino, M. Blonda

ARPA – Agenzia Regionale per la Prevenzione e la Protezione dell'Ambiente, Corso Trieste 27, 70126 Bari, Italia.
*Corresponding author: n.robles@arpa.puglia.it

Abstract:
In recent years, among the renewable energy sources, the photovoltaic plants (PVs) along with wind power plants experienced a remarkable growth rate. If the visual impact assessment of wind turbines is already object of an extensive scientific literature, the effects of PVs on landscape have been few analyzed. A real case-study has is proposed where, with appropriate rendering techniques, a three dimensional simulation of the ground with insertion of PVs pictures has been compared, at different spatial scales, to the bare landscape. In particular, the work is based on the analysis of two parameters affecting negatively the landscape: the visibility and the colour of the plants. The visibility, accompanied by an opportune three-dimensional design stage, and the colour, with the appropriate chromatism of the panels, are mitigation measures for the reduction of the visual impact, and they could play a key role in achieving a virtuous landscape compatibility.
Keywords: Landscape, Landscape Impact, Photovoltaic, Rendering

Introduction

Among the environmental components, the landscape safeguard plays a key role. In the present article landscape analysis is approached considering its scenic structure and sensitivity to transformations, with reference to the visual perception of the landscape fundamental elements, that is the mutual relations between observer's position and some objective characteristics.
Considering the present increasing trend of photovoltaic plants installation, the distribution of the installed power and the number of plants for each region is rather differing; moreover the plant average dimension has increased in almost all regions. It must be underlined that in Northern Italy the average plant dimension is lesser than in the South and, in particular, the largest plants are localized in the Puglia Region. With 19,7% (Fig. 1) this region has the highest national rate of installed registered power, followed by Lombardia with 10,7% and Emilia Romagna with 10,5% (Gestore Servizi Elettrici GSE – Solare Fotovoltaico, Rapporto Statistico 2010).
For this purpose their landscape impact should be frequently assessed. It is then of fundamental importance to assess in advance, from the preliminary planning stage, the impact that this type of plants have on the landscape, intended as common good.
Recently (Torres Sibille et al., 2009) proposed a numerical tool, the OAI_{ssp}, for the objective assessment of the aesthetic impact of solar systems through evaluation of photographic images. In this paper we provide a first step of the OAI_{ssp}, corresponding

to the production of visual simulations of solar plants through 3D_Rendering images of a photovoltaic system of 14 MW in the Apulia Region.

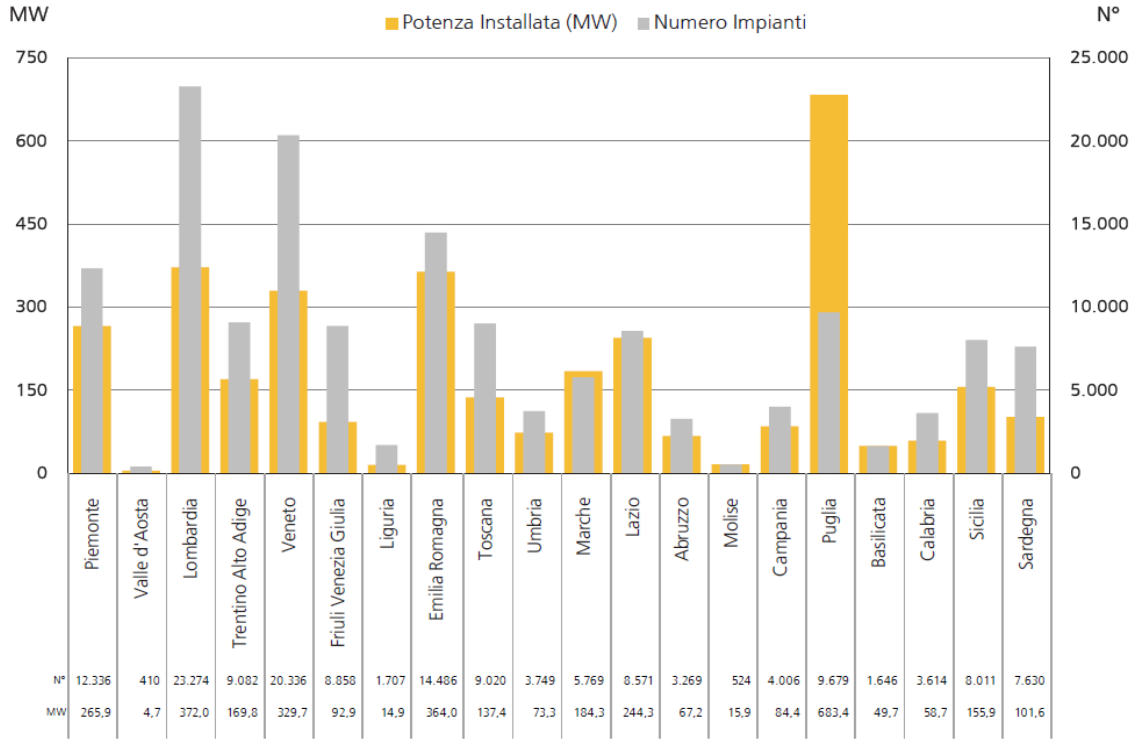


Fig. 1 GSE – Regional distribution of number and power (MW) of photovoltaic plants in Italy (until december 2010)

2. Materials and Methods

A real case-study has been analyzed, specifically a 14 MW photovoltaic plant located in the municipality of Castellaneta (TA) – "Masseria Fresine". It consists of about 60.800 modules rated at 230 Wp, with the capacity to inject an annual amount of 22,34 GWh directly into the grid. The whole plant covers a flat area of 24 hectares in agricultural zone (Fig. 2A).

The type of landscape impact assessment, which uses pictures and environmental insertions (3D_Rendering), falls in the framework of visual simulation techniques for the assessment of landscape compatibility of the projects. In the present case, the impact on the agricultural landscape is concentrated on two important indicators referring to:

- plant visibility;
- plant colour compared to its immediate surroundings.

The degree of perception, depends on these two factors.

Torres Sibille et al. (2009) proposed an objective evaluation of the aesthetic impact of solar plants through evaluation of photographic images. The indicator of the aesthetic impact of a solar panel is expressed through the continuous parameter OAI_{SSP} that falls between 0 and 1. This parameter is a weighted sum of the following aspects:

- the visibility of the plant (sub-parameter I_v);

- the colour of the plant compared to the colour of the immediate surrounding (sub-parameter I_{cl});
- the shape of the plant (sub-parameter I_f);
- the concurrence of various forms and types of panels in the same plant (sub-parameter I_{cc});

where the percentage of each of these sub-indicators on the global indicator value is equal, to 64%, 19%, 9% and 8% rispectively.

3. Results

Fig. 2A shows the plant localization and Fig. 2B highlights what could be observed at a distance of around 300 m from the intervention, close to an overpass of a fast-flowing road. (SS106, E90).

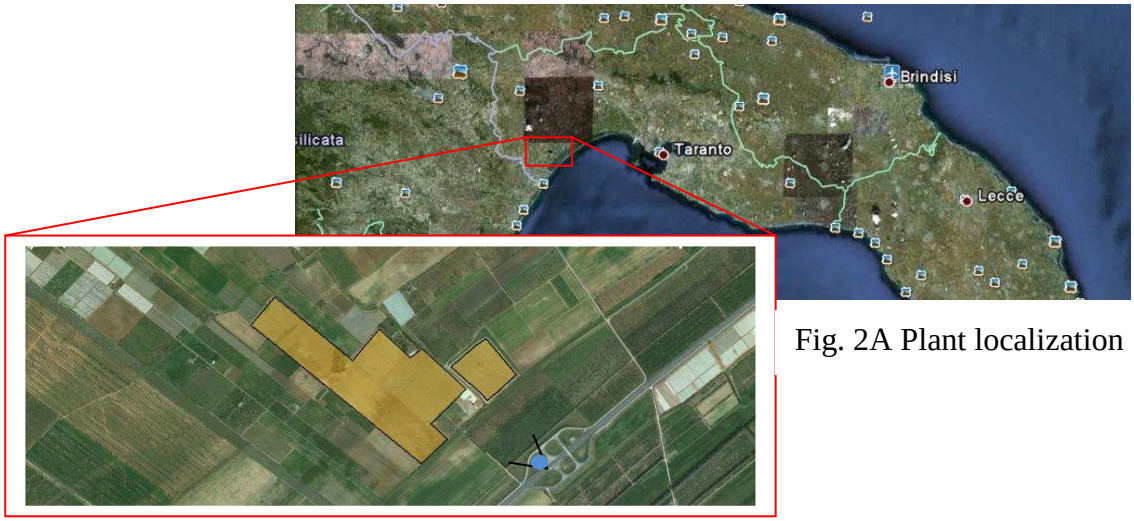


Fig. 2A Plant localization

Fig. 2B Plant and point of view

Fig. 3A shows a view of the actual plant while Fig. 3B shows the same plant with a simulation of chromatic mitigation.

Fig. 3 Plant view (A) and (B) simulation of chromatic mitigation of the photovoltaic plant in Castellaneta (TA) – "Masseria Fresine"

A 3D_Rendering of the color panel without (Fig. 4A) and with (Fig. 4B) a chromatic mitigation has been proposed.

Fig. 4A Rendering 3D with standard color panel

Fig. 4B Rendering 3D with chromatic mitigation

4. Concluding remarks

A correct evaluation of all possible alternatives of shapes, colors, lights, effects of materials and textures should be considered in the design stage of photovoltaic plants for the reduction of their visual impact on the landscape.
In this paper a first step towards an objective evaluation of the visual impact assessment of a photovoltaic plant has been pursued. A complete estimation of the aesthetic quality of the landscape is the result of a complex system (Fig. 5) based on:

- the application of numerical indicators;
- consultation of stakeholders and public opinion:

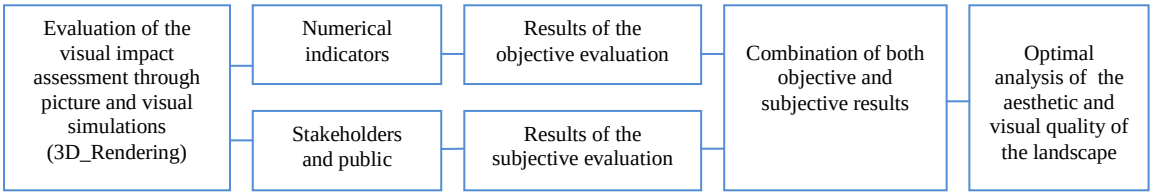


Fig. 5 Flowchart of the methodology for the estimation of the aesthetic quality of the landscape

The use of pictures and visual simulations (Chiabrando et al., 2009), such as the 3D_Rendering technique, represents a useful starting point for the creation of tools necessary for an optimal analysis of the aesthetic and visual quality of the impact of the photovoltaic panels in the landscape.

References

Chiabrando R., Fabrizio E., Garnero G. (2009) La valutazione dell'impatto paesaggistico di impianti fotovoltaici al suolo: proposta metodologica ed esempio di applicazione.

Gestore Servizi Elettrici (GSE) Solare Fotovoltaico, Rapporto Statistico 2010.

Torres Sibille A., Cloquell-Ballester V., Ramirez M. (2009) Aesthetic impact assessment of solar power plants: An objective and subjective approach. Renewable and Sustainable Energy Reviews, 13(5), 986-999.

Marine spatial planning in Apulia (Italy): reconciling seagrass conservation with the multiple use of coastal areas[1]

Simonetta Fraschetti, Giuseppe Lembo, Angelo Tursi, Paolo D'Ambrosio, Antonio Terlizzi, Francesco De Leo, Sidónio Paes, Giuseppe Guarnieri, Stanislao Bevilacqua, Ferdinando Boero

Department of Biological Environmental Science and Technology, Laboratory of Marine Biology, University of Salento, Italy; e-mail: simona.fraschetti@unisalento.it

Abstract: In Apulia (SE Italy) three Marine Protected Areas have been established, but further regulations are needed to guarantee the large-scale protection of critical habitats. The establishment of coherent Natura 2000 networks in Europe, including both terrestrial and marine Sites of Community Importance (SCIs), combined with other conservation strategies, may address several goals that MPAs cannot accomplish alone. In the Mediterranean Sea, *Posidonia oceanica* seagrass meadows are one of the few marine top priorities for the Habitats Directive, deserving protection due to the increasing evidences of fragmentation and loss, despite their key functional role. Along Apulia, current SCIs have been designed on a first seabed mapping carried out at national scale in 1991. In 2006, a new survey allowed map comparison, suggesting evidences of relevant seagrass loss and inadequate SCIs distribution. Here, the use of *a priori* identified conservation targets combined with the analyses of current and emerging human activities offer new scenarios of protection providing a new foundation for ecosystem-based management.

Keywords: Ecosystem-Based Management, Marine Spatial Planning, *Posidonia oceanica*, Apulia, Sites of Community Importance

We acknowledge MIUR (PRIN projects), European Union (SESAME integrated project) and Centro Euro-Mediterraneo per i Cambiamenti Climatici (CMCC) in terms of frameworks for discussions

Regional estimation method of rivers low flow from river basin characteristics

Giuseppe Rossi, Enrica Caporali
Department of Civil and Environmental Engineering, University of Firenze, Firenze, Italy,
giuseppe.rossi@dicea.unifi.it

Abstract: Low flow characteristics are usually estimated from flow gauge stations. However hydrological data are not always available at the site of interest: regional frequency analysis is commonly used for the estimation of flow characteristics at sites where little or no data exists. The study is applied to Tuscany rivers discharge dataset. The area is subdivided into homogeneous regions using an L-moments procedure. The low flow indices Q(7,2) and Q70 at ungauged basins are evaluated with deterministic (Inverse Weighted Distance) and geostatistical (Ordinary Kriging) methods. In order to improve the capability of low flow statistics in ungauged sites a multivariate model, based on geomorphoclimatic characteristics, is also assessed. For each sub-region a relation connecting low flow indices and geomorphoclimatic characteristics is found.

Keywords: regionalization, multivariate analysis, spatial interpolation; L-moments.

1. Introduction

Knowledge of low flow events frequency is required to plan water supply and irrigation systems and moreover to maintain amount and quality of water for wildlife. An estimation of the frequency at which low flow events of different severity might occur is therefore essential for effective water resource planning. Low flow regime is tightly dependent on the catchment hydrogeological feature and a detailed surface and groundwater catchment analysis is necessary for an accurate characterization. However on a practical perspective, although scientifically proven, statistical analysis is often applied to derive indices to characterize low flow regimes and as a measure for low flows. Particularly, low flow frequency behavior is typically characterized using a stochastic approach based on the characterization of some selecting indices (Gustard et al., 1992; Tasker, 1987) thus avoiding to address all the complicated day-to-day variations in the flow record.
Low flow indices can be easily evaluated at gauged sites from observed streamflow time series, but their reliability can be affected by poor and not accurate streamflow data. Sivapalan (2003) indicated that the prediction of surface water flows in ungauged basins is an urgent problem, of immediate relevance to society, dealing with questions such as the impacts of land use and climatic change, biodiversity and sustainable development. In the United States there have been numerous attempts to predict low flows using empirical equations based on catchment area, channel and meteo-climatic characteristics. Another approach to estimate low flow statistics in ungauged sites is the regional statistical analysis, widely used since long time and in different disciplines. It is the most widely used technique in flow estimation in ungauged sites or where few data are available (Riggs, 1973). Regionalization of streamflow characteristics is based on the premise that catchments with similar geology, topography, climate, vegetation, and soils would have similar streamflow responses. It consists of the identification of regional laws, applicable over a more or less wide area, a region, which generally use catchment characteristics as independent variables (Santhi et al., 2008).

2. Materials and Methods

The analysis is carried out on the discharge data recorded in several rivers in the Tuscany Region central Italy by Servizio Idrologico Regionale Toscano (Regional Hydrologic Service of Tuscany) during the period 1949-2008. The main rivers of the region are: Arno, Serchio, and Ombrone Grossetano. Moreover there are small basins of coastal rivers near the Tyrrhenian Sea and the upstream part of Tevere, Fiora and Magra watersheds. A dataset of 65 hydrometric stations is considered, excluding all the discontinuous series, with less than 3 years of data, and stations with long periods of inactivity. The dataset adequately represents the analysed territory.
The area is subdivided into different regions using the L-moments method applied to the 7-day annual minimum flows and to the Q70 annual series. The division into sub-regions was tested using discordancy and heterogeneity statistics (Hosking and Wallis, 1993). A unique region and a subdivision into three different sub-regions, following previous studies on rainfall extremes were considered. Finally a subdivision into five homogeneous sub-regions was undertaken by accounting for hydrological features. With this subdivision the regions are more homogeneous, and the subdivision follows hydrological and precipitation features (Rossi and Caporali, 2010).
An appropriate interpolation technique over the geographical space has to be established in order to determine low flow indices in ungauged sites. For each river section two interpolation techniques, one deterministic (Inverse Weighted Distance) and another one geostatistical (Ordinary Kriging) are applied using the data of the 65 locations of the database.
In order to improve the capability to predict low flow in ungauged sites, a novel multivariate analysis is carried out relating low flow indices and geomorphoclimatic characteristics. The analysis allow to estimate the parameters of a linear correlation between dependent low-flow characteristics and independent catchment and climatic variables. Using a Digital Elevation model (DEM) of the study area, the sub-watersheds for each hydrometric station is found. Each sub-watershed is characterized by means of:

- longest flow paths *FP* [km] (Tucci et al., 1995; Pyrce, 2004);
- topographic mean slope *Sl* [%] (Castellarin et al., 2004; Chokmani and Ouarda, 2004; Laaha and Bloeschl, 2006);
- mean elevation *Hmean* [m a.s.l.] (Gottshalck, 1985; Castellarin et al., 2004; Pyrce, 2004; Laaha and Bloeschl, 2006; Castiglioni et al., 2008; Viglione et al., 2006);
- difference between the maximum and the minimum elevation ΔH [m] (Castellarin et al., 2004; Laaha and Bloeschl, 2006; Vigilance et al., 2006);
- average value of Mean Annual Precipitation *MAP* [mm] (Castellation et al., 2004; Pryce, 2004; Lama and Bloeschl, 2006; Castiglioni et al., 2008; Viglione et al., 2006) available from previous studies (Caporali et al., 2008);
- mean soil permeability *SP* [%] calculated as the percentage of sand into the first 50 cm of the soil (Santhi el al, 2008, Castiglioni et al., 2008). This information is obtained from a pedological map of Tuscany Region cartographic website.

The regionalisation approach requires the development of a regional predictive model for Q70 and Q(7,2). To this aim, the natural logarithms of all geomorphoclimatic characteristics for the 65 sites were regressed against the corresponding Q70 and Q(7,2) values trough a least square mean error procedure. The linear model, used for its simplicity and for the good results it is able to give (Laaha and Bloeschl, 2006), has the form:

$$Q^* = a_1 + a_2 \ln(FP) + a_3 \ln(Sl) + a_4 \ln(Hmean) + a_5 \ln(\Delta H) + {}_6 \ln(MAP) + a_7 \ln(SP) \quad (1)$$

where Q^* is either Q70 or Q(7,2); *FP*, *SL*, *Hmean*, ΔH, *MAP* and *SP* are the explanatory variables of the model, the suitable set of geomorphic and climatic indices; a_i, for i = 0, 1,... , 7, are parameters. The optimal subset of explanatory variables and the estimates of a_i, with i = 0, 1,... , n for both the indices were identified through a least square mean error procedure. Logarithms allow to have variables values easier to be compared (Castellarin et al., 2004) and to have coefficients with a certain homogeneity as well as the same order of magnitude.

3. Results

The procedure is applied to the whole region and then to the other two proposed subdivisions. In Table 1 are summarized the values of the parameters for the different cases. In some subdivisions the equations are reduced eliminating some parameters that show a little correlation with the calculated index.

Index	Subdivision	Sub-region	a_1	a_2	a_3	a_4	a_5	a_6	a_7
Q(7,2)	Unique		-2.02	-5.48	-7.50	10.96	-7.34	1.27	-1.88
Q(7,2)	3 regions	Nord	-19.66	1.64	3.77	4.04		-3.05	1.90
		Centre	-17.80		0.85	2.77	-0.40	0.18	
		South	-5.44	0.13	1.19	0.12	0.69		
Q(7,2)	5 regions	North East	-26.84	3.66	3.07	7.04	-3.50	-4.72	4.93
		North West	-32.27	-3.15	0.98	5.05		-0.58	3.40
		Centre East	6.56	0.60	2.28	0.37		-1.07	-1.32
		Centre West	-21.57		2.33	3.96	-1.91	0.89	-0.38
		South	-5.14	0.19	1.48	0.09	0.60		
Q70	Unique		-19.01	-0.07	1.73	17.85	-4.16	-7.67	-0.63
Q70	3 regions	Nord	-12.53	7.09	7.33	30.67		-20.09	6.99
		Centre	-44.56		2.67	6.58	-1.12	0.90	
		South	-17.78	0.50	3.22	-0.18	2.85		
Q70	5 regions	North East	-22.34	11.34	-3.60	47.20	-0.08	-25.10	8.45
		North West	-80.38	-0.47	0.85	16.21		-6.81	3.14
		Centre East	-33.21	0.45	5.57	4.06		-0.05	0.68
		Centre West	-40.86		7.41	8.65	-5.35	2.26	-1.62
		South	-17.07	0.65	3.88	-0.27	2.66		

Table 1: Parameters of the considered multivariate model.

The models were validated through the calculation of the root mean square error RMSE. The RMSE is calculated for the three proposed subdivisions, for both the proposed low flow indices (Table 2) and, with a jackknife procedure, for the three interpolation methods.
Results for the multivariate analysis confirm the good properties of homogeneity of the final subdivision into 5 regions. For Q70 the RMSE varies from 7.80 with a unique region, to a mean value of 2.89 with the subdivision in three regions to reach a mean value of 1.53 with the subdivision in five regions.
In Figure 1 are shown the results of RMSE for the three interpolation techniques for the two selected indices and the subdivision into 5 regions. Comparing the results of multivariate analysis with the other two interpolation techniques it is possible to state that there is an improving of results especially for the northern regions.

4. Concluding remarks

A method of low flow regionalization is proposed and evaluated. In particular a procedure to evaluate low flow indices in ungauged basins is identified using a regional regression approach. The area is subdivided in 5 regions using an L-moments approach. This subdivision is verified using some interpolation techniques: Inverse Weighted Distance, Ordinary Kriging and a Multivariate Analysis. The results are validated using the jackknife method and calculating the RMSE – Root Mean Square Error for the different subdivisions and the different techniques. The multivariate analysis is the estimation method that performs best. It is able to solve the problems in the two northern regions: in these regions the considered low flows indices present a high variability that can be explained taking into account the geomorphoclimatic characteristics.

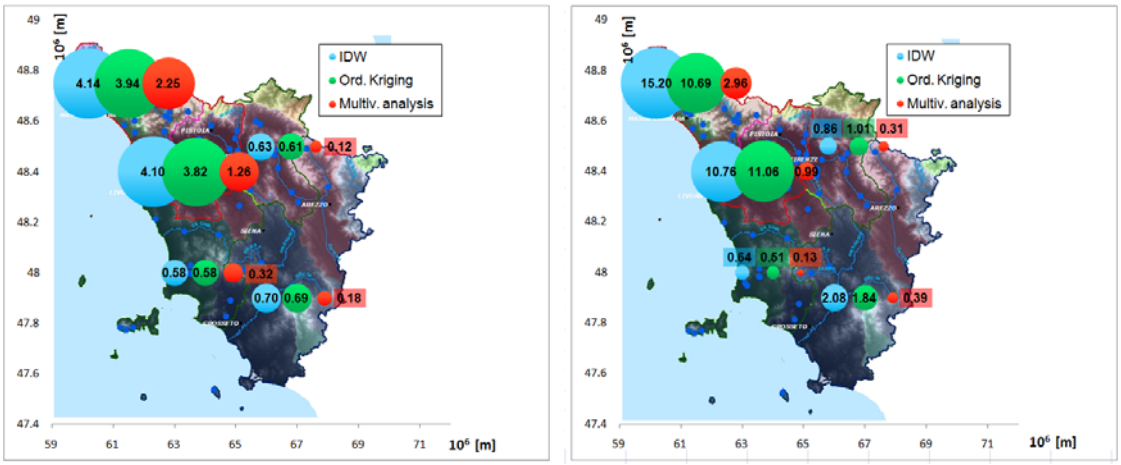


Figure 1: RMSE values for Q(7,2) (left) and Q70 (right) for the subdivision in 5 regions in the three considered interpolation techniques. The circumferences ray is proportional to the RMSE.

References

Caporali E., Cavigli E., Petrucci A. (2008) The index rainfall in the regional frequency analysis of extreme events in Tuscany (Italy), *Environmetrics*, 19, 714-724.

Castellarin A., Galeati G., Brandimarte L., Montanari A., Brath A. (2004) Regional flow-duration curves: reliability for ungauged basins, *Advances in Water Resources,* 27, 953–965.

Castiglioni S., Castellarin A., Montanari A. (2008) Stima delle portate di magra in siti non strumentati mediante tecniche di interpolazione spaziale, *Proc. 31° Convegno Nazionale di Idraulica e Costruzioni Idrauliche*.

Chokmani K., Ouarda T. B. M. J. (2004) Physiographical space-based kriging for regional flood frequency estimation at ungauged sites, *Water Resources Research,* 40, 1–12.

Gottschalk L., Tallaksen L. M., Perzynab G. (1997) Derivation of low flow distribution functions using recession curves, *Journal of Hydrology,* 194, 239–262.

Gustard A., Bullock A., Dixon J. M. (1992) *Low flow estimation in the United Kingdom* (IH Report No. 108). Institute of Hydrology, Wallingford, Oxon.

Laaha G., Bloeschl G. (2006) A comparison of low flow regionalization methods – catchment grouping, *Journal of Hydrology*, 323, 193–214.

Pyrce R. (2004) *Hydrological low flow indices and their uses*, Watershed Science Centre Report No. 04-2004. Trent University, Canada.

Rossi G., Caporali E. (2010). Regional analysis of low flow in Tuscany (Italy), in: *Global Change: Facing Risks and Threats to Water Resources. Friend 2010, Sixth World FRIEND - Flow Regimes from International Experimental and Network Data, Conference*, Servat E., Demuth S., Dezetter A., Daniell T. (Eds.) IAHS Publication 340, 135:141.

Riggs H. C. (1973) *Regional analysis of streamflow characteristics,* US Geological Survey Techniques of Water Resources, United States Government Printing office, Washington.

Santhi C., Allen P. M., Muttiah R. S., Arnold J. G., Tuppad P. (2008) Regional estimation of base flow for the conterminous United States by hydrologic landscape regions, *J. of Hydrology*, 351, 139-153.

Sivapalan M. (2003) Prediction in ungauged basins: a grand challenge for theoretical hydrology. *Hydrological Processes,* 17, 3163-3170.

Tasker G. D., (1987) A comparison of methods for estimating low flow characteristics of streams, *Water Resources Bulletin,* 23, 1077–1083.

Viglione A., Claps P., Laio F. (2006) Utilizzo di criteri di prossimità nell'analisi regionale del deflusso annuo, *Proc. XXX Convegno di Idraulica e Costruzioni Idrauliche.*

Spatial Analysis of some soil physicochemical properties in mountainous massif of Sicó, Portugal [1]

Maria Odete Torres
Centro de Engenharia de Biossistemas, Instituto Superior de Agronomia, Universidade Técnica de Lisboa

Maria Manuela Neves
Centro de Estatística e Aplicações, Universidade de Lisboa and Instituto Superior de Agronomia, Universidade Técnica de Lisboa, manela@isa.utl.pt

Dora Prata Gomes
Centro de Matemática e Aplicações, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Abstract: The mountainous massif of Sicó, with a maximum altitude of 553 meters, is an extensive area of 50.000 ha composed of calcareous Jurassic formations. In these calcareous soils there are some physicochemical properties that are very important to analyze, because of their relevance in the protection of that region. This is a preliminary study where four of those characteristics were measured: soil pH, soil organic matter, plant available phosphorus and soil exchangeable calcium. Classical geostatistical methods are used to analyze separately the spatial variability of the variables under study and to model the dependence structure of the data.

Keywords: soil physicochemical properties, spatial analysis, prediction

1 Introduction

Calcareous soils with high pH present chemical restrictions to support plant growth. These soils have a high capacity to bind phosphorus. On the other hand, hillside soils are subjected to important erosion processes, if not protected by vegetation. The establishment of pastures in these soils can be an important step towards soil protection and to support traditional livestock activity, representing an important element for humanization of the landscapes in less developed regions. Soil organic matter plays an important role in soil quality, productivity and soil resilience to erosion. The variables considered in this study were soil pH, soil organic matter, plant available phosphorus and soil exchangeable calcium. The aim of this study is to

[1] Universidade Técnica de Lisboa and Centro de Matemática e Aplicações, Universidade Nova de Lisboa.

analyze the spatial variability of those soil properties. First an exploratory analysis of the data is made in order to address the spatial variability of those physicochemical properties in the calcareous soils of the massif of Sicó, central Portugal. Using then classical geostatistical methods, Cressie (1993), Goovaerts (1997) and Diggle *et.al.* (1998), the spatial dependency level of those soil attributes is analyzed and their krigged maps are generated. This is the beginning of a more ambitious study that will involve more data collected over different periods. As the region under study is a protected area, to compare the distribution of soil physicochemical properties through several years is another challenge.

2 Materials and Methods

The data underlying this work were collected in that region in October, 1988 at 60 locations where four variables were measured: soil pH, soil organic matter, plant available phosphorus and soil exchangeable calcium.

First an exploratory analysis of the data was performed within the R environment (R Development Core Team, 2006), where many packages are available for the analysis of spatial data. Table 1 shows the summary statistics for the variables under study.

Samples soil properties	Mean	Median	Min	Max	Skewness	Kurtosis
pH	8.185	8.200	8.090	8.500	1.7489	6.7177
P2O5 (mg/ kg)	17.317	17.500	9.000	25.000	-0.3007	-0.6736
Org.Mat.(%)	1.721	1.760	0.950	2.190	-0.4405	-0.4989
Ca (cm(+) /kg)	11.183	11.280	7.990	15.090	0.1865	-0.4217

Table 1: Summary statistics

It can be seen that phosphorus and soil organic matter are skewed; phosphorus presents very high values for skewness and kurtosis coefficients estimates. Soil organic matter reveals a high concentration of high values, see histograms Figure 1.

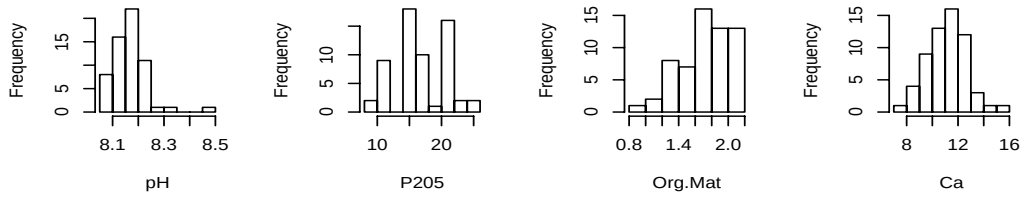


Figure 1: Histograms of the four variables under study

Thus the data were log-transformed to perform the subsequent geostatistical analyses. Sample values of these variables did not show high variability, the highest value was for phosphorus, $CV = 23.1\%$. In accordance with its high skewness, pH showed the presence of outliers, see boxplots Figure 2.

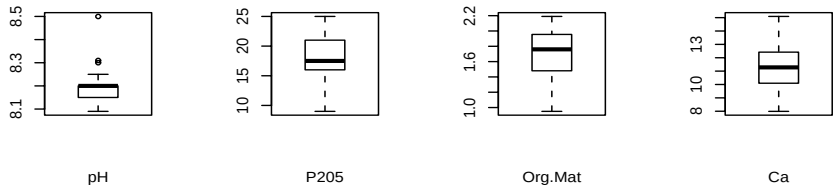


Figure 2: Boxplot of the four variables under study

These first steps in the descriptive analysis led us to construct the experimental semivariograms for lognormal values of pH and soil organic matter. Figure 3 displays the experimental and the theoretical semivariograms and Table 2 gives the parameters of the models fitted to empirical semivariograms.

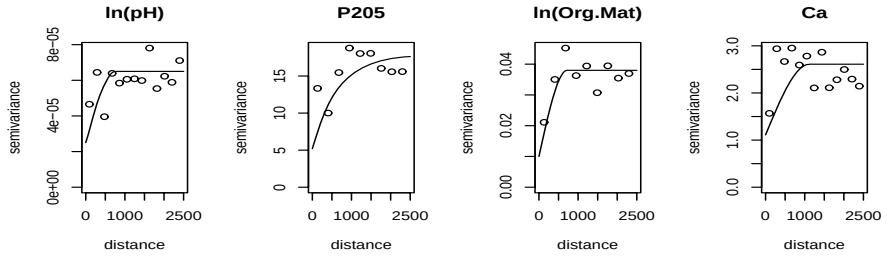


Figure 3: Experimental and theoretical semivariograms for the soil variables

Sample Property	Semivariogram fitted	Nugget	Sill	Range
ln(pH)	Spherical	2.5×10^{-5}	4×10^{-5}	760
P2O5 (mg/ kg)	Exponential	5.2	12.7	650
ln(Org.Mat.(%))	Spherical	0.01	0.028	720
Ca (cm(+) /kg)	Spherical	1.11	1.5	1120

Table 2: Parameters of the fitted models

These first analysis were done using `geoR`, Diggle and Ribeiro (2007) and `gstat`, Pebesma (2011).

3 Concluding remarks

This is a preliminary study using observed values of soil pH, soil organic matter, plant available phosphorus and soil exchangeable calcium, in order to exploit the variability and to look for models for the spatial dependence. Each variable was studied separately and some difficulties were found in fitting models to sampled values. Some more work is needed and is in progress, trying to study the joint behaviour of the variables.

Another challenge topic is to compare distributions of the variables over the years, in order to understand the evolution of the soil in that region, regarding these properties.

References

Cressie N. (1993) *Statistics for Spatial Data.* Revised Edition. John Wiley and Sons. New York.

Diggle P.J., Ribeiro Jr. P.J. (2007) *Model-based Geostatistics.* Springer. New York.

Diggle P.J., Tawn J.A., Moyeed R.A. (1998) Model-based geostatistics. *Applied Statistics* , 473, 299-350.

Pebesma E. J. (2011) The meuse data set: a tutorial for the `gstat` R package. URL *http://cran.r-project.org/web/packages/gstat/vignettes/gstat.pdf.*

Goovaerts P. (1997) *Geostatistics for natural Resouces Evaluation*, Oxford Universiy Press, New York.

R Development Core Team (2006). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Spatial and auto correlation of ecological change: disturbance and perturbation analysis in Circeo National Park (south Latium, Italy).

Gina Galante[1]
1 Sapienza University of Rome – Environmental Biology Department,
gina.galante@uniroma1.it .
Rossana Cotroneo[2]
[2] Statistical National Institute
Stefania Mandrone[3]
[3] Institute for Environmental Protection and Research, ISPRA
Ilaria Strafella[1]
[1] Sapienza University of Rome – Environmental Biology Department

Abstract: "Ecological change" has different meanings: disturbance and perturbation. In this study, disturbance and perturbation were both spatially characterized using Normalize Vegetation Index (NDVI) delta maps (25 years: 1984- 2009) derived by Landsat 5TM imagery. In Circeo National Park, the ecological change spatial pattern was characterized using geostatitsc techniques. Instead, the spatial correlation of data was performed elaborating Euclidean Distance (ED) maps of urban and industrial areas and combining ED maps with disturbance cartography. At 45° a strong anisotropy was revealed by the empirical semivariogram of NDVI losses density, whereas NDVI's gains showed isotropy. The perturbation corresponds to processes of forests re-colonization, whereas the disturbance was human-induced.

Keywords: NDVI, Ecological change, Autocorrelation, Disturbance, Landscape metrics, Spatial correlation

1. Introduction

The ecological systems are heterogeneous, showing a considerable complexity and variability in space and time (Li and Reynolds 1994).Variability and heterogeneity are as well described by all those events that allow modifications or changes (e.g. disturbance) in ecosystem nominal state. Three landscape characteristics may be considered in ecology studies: structure, function and change (Forman and Godron 1986; Gillanders et al. 2008). "Structure" refers to the distribution of energy, materials and species in relation to the sizes, shapes, numbers and types of landscape components. "Function" refers to the interactions between the spatial elements and "change" is usually identified like the "alteration in the structure and function of the ecological mosaic through time". Ecological systems, in fact, are characterized by dynamics, disturbance and change (Reice 1994). The term ecological change can have different ecological meaning: disturbance and perturbation are two aspects of the change. In according with Grime (1979), the removal of biomass from a system constitutes disturbance, or, alternatively, disturbance is a rare and unpredictable event that occurs at different spatial and temporal scale (White 1979, Allen and Star 1982, Rykel 1985; Pickett and White 1985). Remote sensing offers the possibility to identify the reference state of vegetation using an appropriate temporal interval, and the NDVI (Normalized

Difference Vegetation Index). For its character of completeness, it is an excellent carrier of information for both disturbance and perturbation patterns studies (Griffith et al 2002; Zurlini et al 2007). In this study the ecological change phenomenon was detected using the NDVI delta map and change events were disaggregated in its components: NDVI gains (perturbation) and losses (disturbance). The spatial distribution of ecological change and its physical relationships with landscape structure were studied using both geostatistic and spatial analysis technique. The Circeo National Park is a protected area since 1984 and has a superficies of 8.440,00 hectares. This natural reserve is constituted by a variety of different biomass: transitional waters, sandy dunes very rich in alophilic vegetation, Mediterranean and xero-thermophilic forests. Anthropic pressure is relevant in park area due to tourism, intensive agricultural and farming. Disturbance events resulted autocorrelated with a strong anisotropy at 45°, corresponding to the main urban settlements. Perturbation events instead resulted auto-correlated with isotropic pattern, as for a natural driving force. The disturbance data were also spatial correlated with the main urban settlements: disturbance intensity decrease exponentially at the increasing of distance from the main urban areas.

2. Materials and Methods

Two Landsat 5TM remote sensed images with a temporal interval of 25 years (Landsat TM5; July 1984-July 2009, ENVI 3.4) were used to highlight main land use changes in the investigated area. The imagines were acquired respectively in July the 20th 1984 and in July 25th 2009. Imagines were pre-processed to correct atmosphere scattering phenomenon with dark object subtraction method (Chavez 1988, 1996), georeferenced UTM WGS 84 zone 33North, and co-registered to enhance their comparison and superposition (software ENVI 4.7). Band composite (enhancement) and masking with NDVI threshold value techniques were performed to emphasize differences between vegetated and urbanized areas in order to enhance visual interpretation. Principal Component Analysis was applied to the resulting masked vegetated and unvegetated areas to empathise the spectral variance and to better discriminate the different land arrangement within the classes. After the pre-elaboration a supervised classification of land use was performed (Maximum Likelihood categorization algorithm) starting from 30 in field relieved ground true training regions (ROI). In a next step the image difference technique was applied to NDVI maps derived by images using a pixel by pixel's values subtraction (Coppin et al. 2004, Singh 1989): Difference map data were selected for statistical significance by percentiles method. The change thresholds were calculated using the tenth and the ninetieth percentiles of pixels distribution (Fung & LeDrew, 1998) and allows to assign each pixel to one of the following classes: NDVI's increase (perturbation), no change (e.g. stable areas), and NDVI's decrease (disturbance). The output of this procedure is the map of ecological change. Data were exported in GIS environment and study area change map was cut in homogeneous square overlaying a regular grid of 1 km^2. Change density was calculated for any square and geographic coordinate were assigned using the centroid values. Matrix of density values were elaborated for spatial autocorrelation using GS+ software. To asses spatial relationships between disturbance events and anthropic pressure the categorized recent image (2009) was elaborated with spatial analysis ArcGis tool to extract the urban and commercial areas generating two source layers. The Euclidean distance was computed for both this source layers and the resulting features were combined with disturbance

map values. The output matrix explicit how many disturbed pixels were located in the focal classes (source layers) proximity. The data were plotted to highlight spatial relationships between disturbance events and human pressure.

3. Results

The main results can be resumed in two topics: the disturbance autocorrelation data showed a geographical gradient coherent with urban settlements and athropized areas were related with disturbance events. The geostatistical analysis highlighted different patterns of distribution and propagation of ecological change. The disturbance, intended as NDVI losses (biomass losses) semivariogram showed a strong anisotropy in the main urban settlement geographic location. The semivariogram range of about 2 km is coherent with a local scale. Although this is not a cause-effect relation, we can assert that the antrophic areas distribution "justify" the disturbance spatial pattern. This consideration is enforced by the ED results: disturbed pixels decreased exponentially with the urban areas distance increase. On the contrary, perturbation events showed an isotropic spatial autocorrelation: biomass increase proceeded with a natural pattern of distribution and is not influenced by urban and commercial areas location. Moreover, NDVI gains resulted cross-correlated with stable areas, as for an enhancing effect of stability on biomass increase. The perturbation events were also related with urban and commercial ED maps: NDVI gain pixels increase with an opposite trend on respect the NDVI losses.

4. Concluding remarks

The natural pattern of NDVI gains distribution is drive by natural forcing, in fact, the Tobler axiom states that "The closest things are more similar than those distance", and this is the true variables nature of ecological data. In particular, a fundamental ecological process as biomass primary production tends to be constant at high hierarchical level (landscape level). Instead, NDVI losses are not always spontaneous (as for fire, storms or spontaneous vegetation regression dynamics), but often caused by human activity. In this case, we appreciate a non-natural forcing that modifies the distribution pattern of change: gradients are evident and coherent with anthropized area distribution. It is important to establish directional pattern of ecological change to modify the environmental policies of natural and protected areas management and to enhance the natural resources recovery.

References

Allen TFH, Star TB, (1982) Hierarchy, Perspectives for Ecological Complexity. Chicago Press, Chicago, p. 310.

Chavez PS Jr, 1996 Image-Based Atmospheric Corrections-Revisited and Improved. Photogramm. Eng. Rem. S. 62(9), 1025-1036.

Coppin P., Jonckheere I., Nackaerts K., Muys B. and Lambin E. (2004) Digital change detection methods in natural ecosystem monitoring: a review. International Journal of Remote Sensing, 25:9, 1565-1596

Forman, R.T.T., Godron, M. 1986. Landscape ecology. Wiley, New York.

Fung T, LeDrew E (1988) The determination of optimal threshold levels for change detection using various accuracy indices. Photogramm. Eng. Rem. S. 54(10), 1449-1454.

Gillanders SN, Coops NC,Wulder MA, Gergel SE, Nelson T (2008) Multitemporal remote sensing of landscape dynamics and pattern change: describing natural and anthropogenic trends. Prog. Phys. Geog. 32(5), 503-528.

Griffith J. A., Martinko E. A., Whistler J. L., Price K. P. (2002) Preliminary Comparison of Landscape Pattern-Normalized Difference Vegetation Index (NDVI), Relationships to Central Plains Stream Conditions. Journal of Environmental Quality, 31:846-859.

Li H., Reynolds JF. (1994) A simulation experiment to quantify spatial heterogeneity in categorical maps. Ecology, 75:2446–55.

Pickett STA, White PS (1985) The ecology of natural disturbance and patch dynamics. Academic Press, Orlando

Reice SR (1994) Nonequilibrium determinants of biological community structure. American Scientist, 82, 424-35.

White PS (1979) Pattern, process and natural disturbance in vegetation. Bot. Rev. 42, 229-299.

Singh A. (1989) Digital change detection techniques using Remotely sensed Data. Remote Sensing and Tropical Land Management, Eden M.J. and Parry J.T. (eds), John Wiley & Sons, London, pp. 237-254.

Zurlini G, Riitters KH, Zaccarelli N, Petrosillo I (2007) Patterns of disturbance at multiple scales in real and simulated landscapes. Landscape Ecol. 22, 705-721.

Spatial diversity in a "zoom-lens": Analysing ecological communities through weighted spatial scales [1]

A.C. Studeny, C. Brown, J.B. Illian

Centre for Research into Ecological and Environmental Modelling
University of St Andrews, St Andrews, UK
Email: angelika@mcs.st-and.ac.uk

Abstract: Traditional summary statistics for biological diversity are rarely able to fully identify processes maintaining the biodiversity of ecosystems. However, these processes produce patterns *in space*, and spatial statistics offer a way to describe such signals and so identify underlying mechanisms.
A new community-level measure of spatial structure, the cross-pair overlap distribution (xPOD), summarises the spatial overlap of species pairs over a scale defined by radius R. Here, we extend this approach by developing a radius-weighted version of the xPOD which provides greater flexibility over spatial resolution. In particular, this allows us to identify behaviour related to different mechanisms, which can then be linked to the spatial scale at which they operate. In its general form, this approach can be applied to any community-level second-order summary statistic which is a function of scale, leading to a considerable increase in informative power.

Keywords: spatial point pattern, second-order characteristics, biodiversity, pair-correlation function, cross-pair overlap distribution

1 Introduction

A central problem in community ecology is to link ecological processes to observed patterns [12, 1]. Traditionally, a focus has been on the species abundance distribution as a diagnostic for community-scale effects [3, 4, 11, 8],

Summary statistics based on the species abundance distribution (*first-order characteristics*) are also used to quantify biodiversity, which is typically measured in non-spatial terms [7]. Recently, it has been shown that the species abundance distribution does not hold enough information on community structure in (highly diverse) tropical rainforests to draw inferences about underlying processes [2, 10]. However, spatial properties are found to be informative as they can capture the interactions that structure the community at the local scale [6, 9]. They are summarised by *second-order characteristics* in spatial point process analysis.

[1] ACS is funded by the University of St Andrews, CB is funded through a Microsoft scholarship

The cross-pair overlap distribution (xPOD) provides a new community-level measure of spatial diversity [2]. First applications have demonstrated its ability to reliably distinguish between different processes generating community patterns. A short-coming of the xPOD as well as other second-order characteristics is their rigidity with respect to the spatial scale R. Though not restricted to a specific value, this must be set before the statistic is evaluated. Scales at which specific processes operate are largely unknown [13] and information can be lost by an arbitrary choice of evaluation radius R as a result.

This paper introduces a radius-weighted version of the xPOD. It is no longer a function of R and hence increases the flexibility over spatial resolution. This approach is easily generalised to any community-level second-order summary statistic which is a function of scale. It is not limited to application in ecology, but extends methods for multi-type point pattern analysis in general which are relevant in many different fields of study.

2 Materials and Methods

Consider a multi-type (marked) point process $M = [(x_n, y_n); m(x_n, y_n)]$ where the number of types $m(\cdot) \in \{1, ..., S\}$ is finite, on a window of unit area ($x., y. \in [0, 1]$). This can, for example, describe the positions of tree species in a sample area. (For details on point processes and their statistics, in particular the cross-pair correlation function, see [5]).

The cross-pair overlap distribution

We calculate the xPOD as a description of interspecific clustering within M. Based on the cross-pair correlation function g and a chosen spatial scale R (typically $R \ll 1$), it gives the distribution across all types i, j of

$$A_{ij}(R) = \int_0^R \log g_{ij}(r) dr. \tag{1}$$

A radius-weighted version of the xPOD

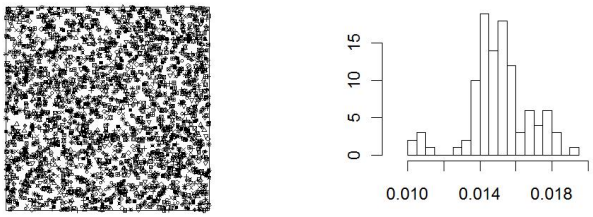
Eqn. (1) can be interpreted as the (rescaled) expectation of $\log g$ over the spatial scale r where equal weights are placed on $[0, R]$:

$$\frac{1}{R} A_{ij} = \int_0^1 \log g_{ij}(r) f(r) dr \tag{2}$$

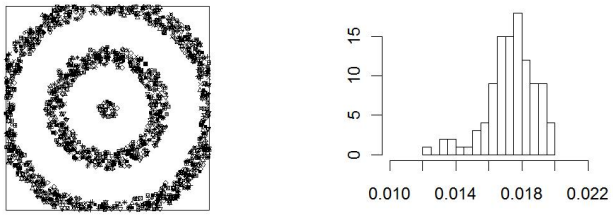
where $f(r) = \frac{1}{R}\mathbf{1}_{[0,R]}$ is the uniform distribution. By choosing a different probability distribution function (pdf) f we can introduce non-uniform weights on the spatial scale, e.g.

$$\tilde{A}_{ij} = \mathbf{E}[\log g] = \int_0^1 \log g(r) Beta(r; \alpha, \beta) dr. \tag{3}$$

Figure 1: Two point patterns and their unweighted xPODs



(a) Random point pattern

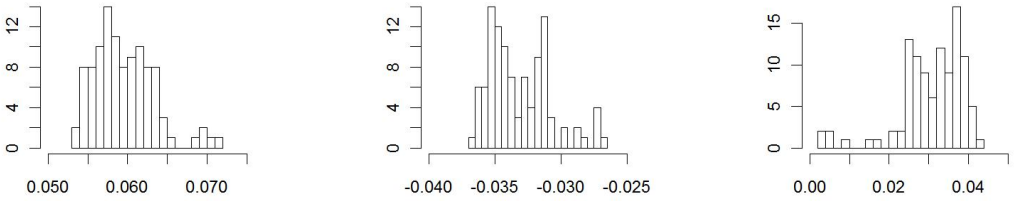


(b) Structured point pattern

The *Beta* distribution was chosen for its easy interpretation as weights as well as for its natural range of $[0, 1]$. However, any other pdf for which the expectation in (3) exists can be used instead. The parameters of the *Beta* distribution determine the focus on certain spatial scales (local neighbourhood, intermediate distance, far distance, or combinations of these) while considering the whole point pattern. Crucially, this generalization no longer requires the evaluation scale R to be set in advance.

We demonstrate the performance of the radius-weighted version of the xPOD given in eqn.(3) by means of two very different point patterns - a random Poisson point process and a highly structured pattern (Fig.1). The random pattern provides the usual reference point, while the highly structured pattern was chosen to show a similar shape of the basic xPOD. The parameters for the *Beta* distribution are set to (1) $\alpha_1 = 1, \beta_1 = 3$, (2) $\alpha_2 = 3, \beta_2 = 3$ and (3) $\alpha_3 = 3, \beta_3 = 1$. This corresponds to zooming in on local, intermediate, and large-scale behaviour. We expect the differences between the patterns to produce divergent xPODs at different spatial scales.

Figure 2: Weighted xPODs for the structured point pattern



(a) $Beta(1, 3)$ (b) $Beta(3, 3)$ (c) $Beta(3, 1)$

3 Results

In their original (unweighted) version, the xPODs of the two point patterns are indistinguishable (Fig.1) as the xPOD cannot express scale-specific behaviour.

When the radius-weighted version is applied, the xPOD for the Poisson pattern remains virtually the same, independent of the chosen weighting – as we expect given the self-similarity of the point process across all scales. For the structured point process, on the other hand, the change in structure with scale is now clearly visible (Fig.2): At small radii, marks or 'species', occur together and consistently overlap more than they would if the entire pattern was random. Hence, when weighted towards very local behaviour, the xPOD is centered around positive values. At medium radii this behaviour changes. Marks can now be wholly separated from one another by the empty areas between rings. This leads to a change in sign when the xPOD is focussed on this scale. At large radii, neighbouring rings in the pattern are encountered, and so the values in the distribution become positive again.

4 Concluding remarks

Traditional spatial summary statistics are often not flexible enough to distinguish processes that operate on different spatial scales. Extending one of these statistics, the cross-pair overlap distribution, to radius-weighted version, we were able to discriminate patterns while first- and traditional second-order characteristics conceal their radically different behaviour. This suggests that spatial statistics may be adapted in a way that takes advantage of issues of scale in order to increase their explanatory power. This is relevant to ecology, where important processes are scale-dependent, and where the ability of spatial statistics to reveal such processes is of increasing interest.

References

[1] Bolker, B., Pacala, S.W. (1997) Using moment equations to understand stochastically driven spatial pattern formation in ecological systems. *Theoretical Population Biology*, 52, 179-197.

[2] Brown, C., Law, R., Illian, J., Burslem, D. (2011) Linking ecological processes with spatial and non-spatial patterns in plant communities. *In review.*

[3] Hubbell, S.P. (2001) *The unified neutral theory of biodiversity and biogeography*, Princeton University Press, Princeton, NJ.

[4] Hurlbert, A.H. (2004) Species-energy relationships and habitat complexity in bird communities. *Ecology Letters*, 7, 714-720.

[5] Illian J.B., Penttinen A.,Stoyan H., Stoyan D. (2009) *Statistical Analysis and Aodelling of Spatial Point Patterns*, Wiley, New York.

[6] Law, R., Illian, J.B., Burslem, D.F.R.P., Gratzer, G., Gunatilleke, C.V.S. & Gunatilleke, I.A.U.N. (2009) Ecological information from spatial patterns of plants: Insights from point process theory. *Journal of Ecology*, 97, 616-628.

[7] Magurran, A.E. (2004) *Measuring Biological Diversity*, Blackwell Science, Oxford.

[8] McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., Dornelas, M., Enquist, B.J., Green, J.L., He, F., Hurlbert, A.H., Magurran, A.E., Marquet, P.A., Maurer, B.A., Ostling, A., Soykan, C.U., Ugland, K.I., White, E.P. (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, 10, 995-1015.

[9] McGill, B.J. (2011) Measuring the spatial structure of biodiversity. pp.152-171, in: *Biological Diversity - frontiers in measurement and assessment*, Magurran, A.E. & McGill, B.J. (Eds.), Oxford University Press, Oxford.

[10] Rajala, T.A., Illian, J.B. (2011) A family of spatial biodiversity measures based on graphs. *Environmental and Ecologial Statistics. In submission.*

[11] Thibault, K.M., White, E.P., Ernest, S.K.M. (2004) Temporal dynamics in the structure and composition of a desert rodent community. *Ecology*, 85, 2649-2655.

[12] Watt, A.S. (1947) Pattern and process in the plant community. *Journal of Ecology*, 35, 1-22.

[13] Wiegand, T., Gunatilleke, S., Gunatilleke, S. (2007) Species associations in a hetergeneous Sri Lankan dipterocarp forest. *The American Naturalist*, 170, E77-E95.

Spatio-temporal changes of biodiversity indices in the bathyal demersal assemblages of the Ionian Sea

Maiorano P., Giove A., Minerva M., Sion L., D'Onghia G.
Department of Biology, University of Bari "Aldo Moro"
p.maiorano@biologia.uniba.it

Pollice A., Ribecco N., Muschitiello C.
Department of Statistics "Carlo Cecchi", University of Bari "Aldo Moro"

Abstract: Spatio-temporal analysis of biodiversity indices estimated in the bathyal demersal species assemblages of the Ionian Sea has been performed. Data were collected during 16 trawl surveys carried out from 1995 to 2010 as part of the international MEDITS project funded by EC. In the Apulian sector a significant increase of species richness and a significant decrease of evenness have been detected. In the Southern Calabrian sector a significant decrease of evenness has been detected while a positive trend has been found for the Simpson index. GAM's have then been applied to explain the dependence of the indices in terms of time and space.

Keywords: Spatio-temporal analysis, biodiversity, deep-sea assemblages, Generalized Additive Models, Mediterranean.

1. Introduction

Loss of biodiversity, and its possible consequences on both ecosystem functioning and services, has promoted the assessment and monitoring of species diversity in marine ecosystems. Research on the deep-sea assemblages has been also carried out in the Ionian Sea during the last years (D'Onghia et al., 1998, 2003, 2004). The previous studies indicate depth and geographical area as the main factors influencing the fauna assemblages of the Ionian Sea. In this context, a spatio-temporal analysis of the biodiversity indices estimated in the bathyal demersal species assemblages of the Ionian Sea has been performed in order to monitor the diversity in the bathyal demersal assemblages in two geographic areas of the Ionian Sea.

2. Materials and Methods

Data were collected during 16 trawl surveys carried out from 1995 to 2010 as part of the international MEDITS project funded by EC (Bertrand et al., 2000). The samples analyzed come from a total of 260 and 236 hauls carried out in the Apulian and Southern Calabrian sectors respectively, between 200 and 800 m in the spring season (May-June). In each sector, the biodiversity indices of Margalef (species richness, $\boldsymbol{d}$),

	Species richness (d)		***Pielou's evenness (J)***		***Simpson index (λ)***	
parameters	*estimate*	*p-value*	*estimate*	*p-value*	*estimate*	*p-value*
intercept	2.754	<0.000	0.651	<0.000	4.637	<0.000
smooth terms	*df*	*p-value*	*df*	*p-value*	*df*	*p-value*
s(lon, lat)	24.108	<0.000	22.650	<0.000	22.830	0.006
s(year)	6.863	<0.000	3.155	<0.000		

Table 1: GAM's estimates for the indexes ***d***, ***J*** and ***λ***: estimates of the parametric terms, and degrees of freedom of the non-parametric ones with respective p-values.

Pielou (evenness, ***J***) and Simpson (***λ***) were estimated using density data (N/km^2) obtained for a total of 117 and 110 species collected in the demersal assemblages of the Apulian and Southern Calabrian sectors respectively. Data analysis was carried out as part of OBAMA project funded by MIUR. The trawl geographic coordinates were considered as the main spatial information contained in the data. Such information is redundant with the sector specification and the depth measurement. GAMs were applied to analyze the indices variation in terms of time and space. The choice between competitive models, characterized by different combinations of response distributions, link functions and predictors, was performed in terms of effects significance and overall model fit.

3. Results

While ***d*** and ***J*** could be considered Gaussian in the GAM specification, the Gamma assumption allowed to account for the asymmetry observed in the empirical distribution of the Simpson index. Table 1 reports the results for the models chosen to represent the three biodiversity indices. Species richness shows significant nonlinearity of both spatial and temporal effects.
As can be seen in Fig. 1 (a) the species richness index shows an increasing temporal evolution with a peak around 1998.
Level curves in Figure 2 (a) represent the spatial behaviour of this index confirming smaller values in deeper waters. Time has a slightly negative effect on Pielou's evenness (Fig. 1, b) and the nonlinearity of the spatial component appears to be strongly significant. Level curves in Figure 2 (b) represent the spatial behaviour of this index with values increasing with depth.
Time does not have a significant impact on the Simpson index and the spatial behaviour shows some peculiarities with respect to the other two, with different evidence between areas.

4. Concluding remarks

The present study shows changing values in the species richness related to time and geographic area. The spatial pattern confirms a decreasing diversity with depth reported in previous studies (D'Onghia et al., 2003, 2004). The increase in the species diversity

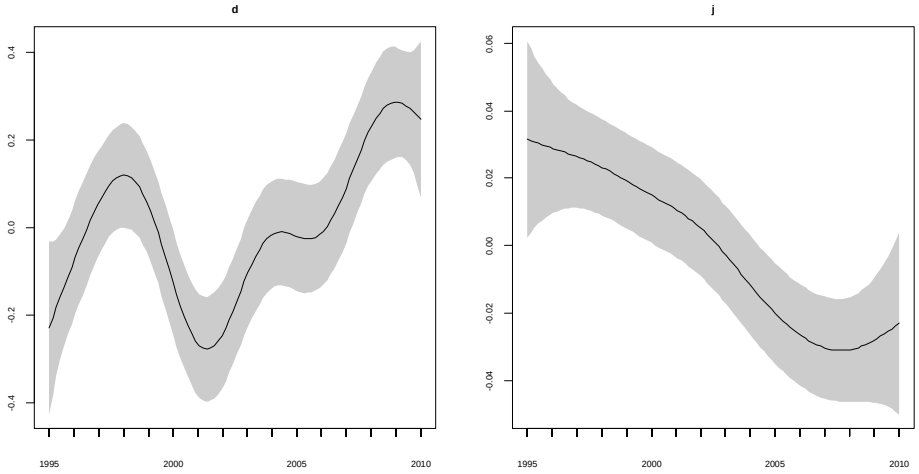


Figure 1: GAM estimates of the time effect for the species richness ***d*** and the Pielou's evenness ***J***.

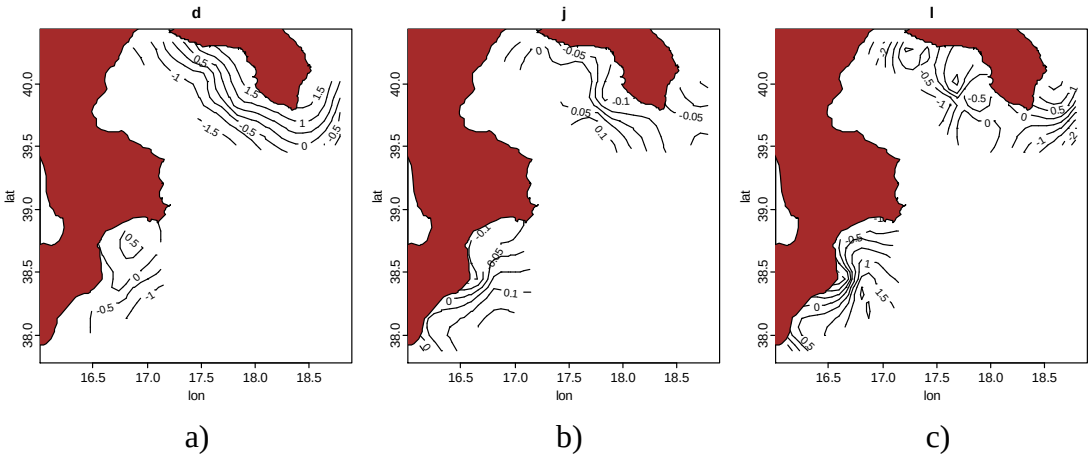


Figure 2: Gam estimates of the spatial effect for the indexes a) ***d*** (richness), b) ***J*** (evenness) and c) lambda (Simpson Index).

with time can be related to the new species recorded in the last years as already reported in the Ionian Sea (Maiorano et al., 2010). The variation observed in the evenness index does not reveal a significant difference between the two geographic areas while a different pattern was observed for the Simpson index mostly probably due to the environmental conditions and fishing pressure between Apulian and Calabrian areas (Capezzuto et al., 2010).

References

Bertrand J.A., Gil de Sola L., Papaconstantinou C., Relini G., and Souplet A., (2000) An international bottom trawl survey in the Mediterranean: the MEDITS programme, IFREMER *Actes de Colloques* 26: 76-93.

Capezzuto F., Carlucci F., Maiorano P., Sion L., Battista D., Giove A., Indennidate A., Tursi A., D'Onghia G., 2010. The bathyal benthopelagic fauna in the NW Ionian Sea: structure, patterns and interactions. *Chemistry & Ecology*. Volume 26 Supplement 1: 199-218.

D'Onghia G., Mastrototaro F., Matarrese A., Politou C.Y., Mytilineou C. (2003). Biodiversity of the upper slope demersal community in the Eastern Mediterranean: preliminary comparison between two areas with and without trawl fishing. *Journal of Northwest Atlantic Fishery Science*, 31: 263-273.

D'Onghia G., Politou C. Y., Bozzano A., Lloris D., Rotllant G., Sion L., Mastrototaro F. Pp. (2004). Deep-water fish assemblages in the Mediterranean Sea. *Scientia Marina*, vol. 68 (SUPPL. 3); p. 87-99, ISSN: 0214-8358.

Maiorano P., Sion L., Carlucci R., Capezzuto F., Giove A., Costantino G., Panza M., D'Onghia G., Tursi A., 2010. The demersal faunal assemblage of the North-Western Ionian Sea (Central Mediterranean): present knowledge and perspectives. *Chemistry & Ecology*. Volume 26 Supplement 1: 219-240.

Wood S. N. (2006) *Generalized Additive Models: An introduction with R*, Chapman & Hall/CRC.

Spatio-temporal variability in stream flow status: Candelaro river case study

A.M. De Girolamo, A. Calabrese, G. Pappagallo, G. Santese, A. Lo Porto
Water Research Institute, National Research Council of Italy
annamaria.degirolamo@ba.irsa.cnr.it

Francesc Gallart
Institute of Environmental Assessment and Water Research (IDÆA), CSIC

Narcis Prat
Dept. Ecologia, Univ. Barcelona

Jochen Froebrich
Centre for Water and Climate (CWK) Integrated Water Resources Management
Wageningen UR - Alterra

Abstract: The spatio-temporal variability of the flow status of the Candelaro river and its tributaries was assessed in order to analyze the hydrological regime and to provide some information to assist the determination of the ecological quality as required by the European Water Framework Directive. Different types of flow were defined for this temporary river and a flow status frequency method was used to analyse the occurrence along the year of the different flow conditions. Daily streamflow data recorded in some gauging stations were used and the results were verified through field observations of flow status. Based on the results, maps were developed representing the spatial distribution of the different flow types along the year.
Keywords: Temporary river, Water Framework Directive, Hydrological regime.

1. Introduction

The European Water Framework Directive (WFD) constitutes a new view of water resources management in Europe, based mainly upon ecological elements, its final objective is achieving at least “good chemical and ecological quality status” of water bodies. To attain good ecological status, aquatic systems must not significant depart from reference “natural” conditions. Information describing stream hydrological regime are of the major importance for implementation of WFD, since it may be responsible for the ecological status. The analysis of the hydrological regime is particularly relevant for the intermittent rivers since it varies on spatial and temporal scale depending on precipitation patterns and is severely disturbed by flash floods. Many definitions of non permanent rivers can be found in the literature (Svec et al., 2005) and in the WFD implementation process some EU countries have developed a definition of these water bodies based on the number of days per year during which water is flowing in the river. In 2008, Italy differs these stream types, and in 2009-2010 gives the criteria for the monitoring activities, but it doesn’t include the timing of samplings. Frequency and

timing of samplings are crucial points for temporary rivers because the quantitative flow status determines the river biological communities. Streams with long dry periods can have a reduced fauna compared with permanent but their ecological status can be "good" even if the flow is scarce and only pools remain along the river network. In this framework, the main objective of the present paper is to analyse the spatio-temporal variability of the flow status of the Candelaro river and its tributaries in order to classify the hydrological regime and to provide some information to assist the determination of the water ecological quality.

2. Materials and Methods

2.1 Study area

The Candelaro river basin is located in the Puglia region in southern Italy (Figure 1). The basin is characterised by a mean elevation of 300 m above sea level, ranging from 0 m to 1142 m. The drainage area is about 2200 km^2 and the main river course has a length of 67 km. The soils are related to the lithology and generally show a texture varying from sandy-clay-loam to clay-loam or clay. The average annual precipitation in the catchment in the period from 1986 to 2001 was 579 mm. The rainfall is mostly concentrated in autumn and winter, it is unevenly distributed and often occurs with high intensities of short duration. The stream flow regime changes rapidly and follows the precipitation regime closely.

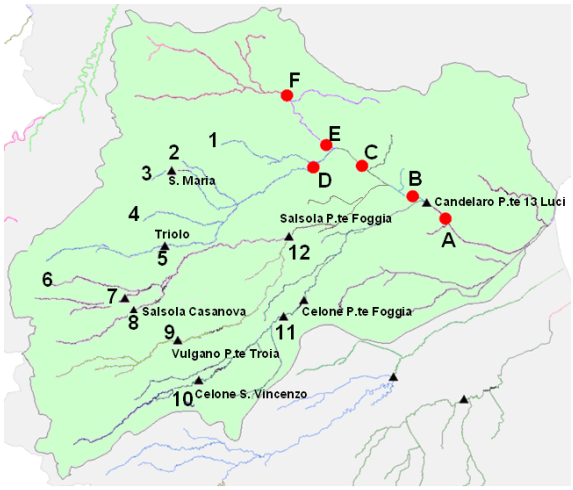


Figure 1: Candelaro river basin. River bodies identified by Puglia Region, streamflow gauging stations (triangles), simulated streamflow (red points).

2.2 Methods

In 2010, the Puglia region authority provided a river characterization based on the use of abiotic indicators, following the System B of the WFD and the national Decree 131/2008. Seven river types were identified in the Candelaro river basin which in turn were differentiated in 14 river bodies according to anthropogenic pressures.

For each river body we have analyzed the streamflow at reach scale as proposed within the MIRAGE Project[1] (Gallart et al., 2010). Five classes of flow were identified as relevant to aquatic life: *dry* (if most of the reach is dry), *disconnected pools* (when the flow is scarce only isolated pools appear along the channel), *connected pools* (when there are a lot of pools connected by a slow flow), *riffles* (if the flow is continuous) and *floods*. The monthly frequency of occurrence of these flow statuses were evaluated. In order to do this, it was necessary to fix the thresholds of streamflow between one class and another. At a first analysis these flow values can be determined on the basis of the Flow Duration Curves (FDC), but field observation are necessary to verify the thresholds between disconnected pools and connected pools and between connected pools and riffles. The monthly frequency of each flow status was calculated on measured monthly flows, when available (from 1965 to 1996), or on simulated values. The hydrological model "SWAT" was used to simulate streamflows from 1990 to 2009.

3. Results

Figure 2 shows the FDC (a) and the frequency of occurrence of the five streamflow classes (b) of the Celone river (gauge 10 in Figure 1). It shows a seasonal hydrological regime. Here, dry conditions generally occur from May to December; floods are frequent from January to March, and disconnected pools can take place from April to December.

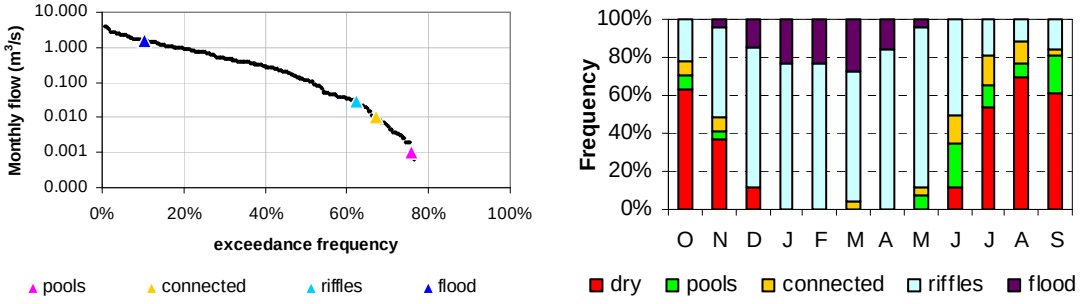


Figure 2: (a) Flow Duration Curve of the Celone S. Vincenzo gauging station (85 km^2). The triangles represent the thresholds identified between two different statuses. (b) Flow status frequency graph of the Celone river.

All the streams located in the mountainous part of the basin are quite natural; they show the same streamflow regime (gauging stations N. 6, 7, 8, 9, 10, in Figure 1). In this area biological samplings should be scheduled before June because after this month dry conditions can take over.

On the ground level of the basin the situation is different. The few natural reaches (1, 2, 3, 4 in figure 1) can be affected by dry conditions, as well as by isolated pools, also in the winter season. On the contrary, the main channel, which is heavily modified, shows a regime very far from its natural conditions (gauging stations A, B, C, D, F in Figure 1). Many waste water treatment plants discharge their sewages into the river, consequently the low flow regime is completely altered and it never reaches dry

[1] European Community's Seventh Framework Programme (FP7/2007-2011). MIRAGE Project (211735)

conditions. For the same reason, the downstream reaches of Salsola river show a dry status period shorter than that recorded in headwater streams. The Celone river, which is one of the main tributaries of the Candelaro, shows in its downstream reaches a very long dry period, from April to February. This is mainly due to a reservoir, built in 1996, that alters the natural regime of the river.
In 2010, monitoring activities were carried out along all over the river network. During the wettest period (January) continuous flow was recorded all over the streams, while during the driest period (September) only the 7% of the river network presented a continuous flow. Figure 3 shows a map representing the flow statuses during the driest period.

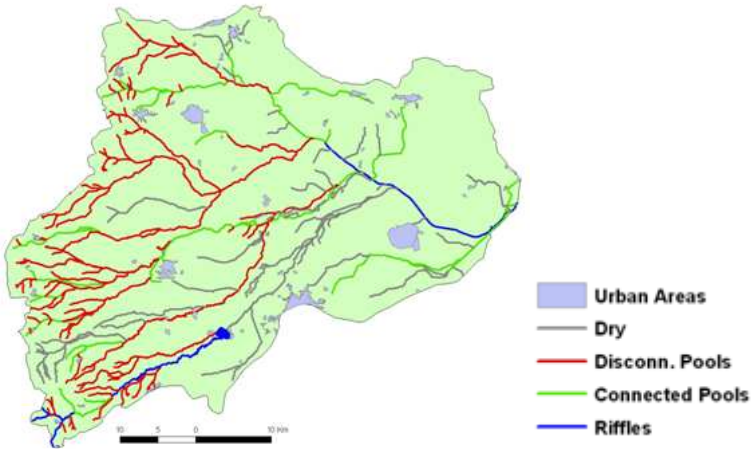


Figure 3: Flow status recorded during the driest period in the Candelaro river basin.

4. Concluding remarks

Flows are highly variable both in space and time in the Candelaro river basin. Dry and disconnected pools statuses are very frequent and their duration varies both year to year and from reach to reach. Hence, biological samplings have to be scheduled taking into account the flow statuses at reach scale. A new method has been used to describe streamflow regime at reach scale. The results achieved show that the flow status frequency graphs can really provide useful information in order to evaluate ecological water quality in temporary rivers.

References

Gallart, F., Prat N. et al., (2011 in revision). Analyzing stream reach flow status frequency to assist the determination of ecological quality in temporary streams, Aquatic Sciences.

Svec, J.R., Kolka, R.K., Stringer, J.W. (2005). Defining perennial, intermittent, and ephemeral channels in Eastern Kentucky: Application to Forestry best management practices. Forest Ecology and Management, 214. 170-182.

Statistical assessment of the plant protection level within protected areas (PA) based on remote sensing products

Maria Elena Menconi, Ciro Luca Pacicco
"Uomo e Territorio" Department, University of Perugia, Italy
melenamenconi@yahoo.it

Abstract:

The study tested whether long established protected areas (PA) in central Italy are more effective in protecting vegetation, than areas with the same characteristics but no formal protection. The index used for the comparison is the EVI (Enhanced Vegetation Index), measured by the Terra satellite with the MODIS instrument. For this index has been evaluated the spatial variability of the highest values and their correlation with land use, geographic and morphological characteristics of study area using the frequency ratio (FR) method. The work methodology applied to the regions of central Italy showed that, on equal terms, the areas within PA present higher values of the EVI index compared to the remaining areas.

Keywords: remote sensing, EVI index, protected areas, frequency ratio

Statistical calibration of the Carlit index in the Pontine Island of Zannone

Giovanna Jona Lasinio, Maria Alessandra Tullio
Department of Statistical Sciences, "Sapienza" University of Rome,
giovanna.jonalasinio@uniroma1.it

Nadia Abdelahad, Edoardo Scepi, Simona Sirago
Environmental Biology Department, "Sapienza" University of Rome

Alessio Pollice
Department of Statistical Sciences "Carlo Cecchi", University of Bari

Abstract: The WFD[1], adopted by the European Community requires that Member States achieve and maintain a good ecological status of all water bodies by 2015. In the marine context, the ecological status has to be quantified applying indexes based on appropriate key biological elements. The CARLIT index is a cartographic monitoring tool enabling the EQR[2] to be calculated using macroalgae in coastal hard bottoms as a key biological element. Here we investigate the role of *Cystoseira amentacea var.stricta:* a key macroalgae involved in the index definition. We analyze the relation between the algae presence and geomorphological characteristics of Pontine Islands coast through standard logistic regression and autologistc models to account for spatial correlation.

Keywords: bioindicator, logistic regression, autologistc model

1 Introduction

The Water Framework Directive (WFD) 2000 /60 /EC, adopted by the European Community in 2000 requires that Member States achieve and maintain a good ecological status of all water bodies by 2015. In the marine context, the ecological status has to be quantified applying indexes based on appropriate key biological elements, which allow the categorization of water bodies into five Ecological Status (ES) classes. In order to implement the WFD, several indices based on macroalgae have been proposed. One of them is the CARLIT index (Ballesteros et al. 2007), which has been adopted for the evaluation of Italian rocky coasts. A recent application of the CARLIT protocol to the entire coast of the five Pontine Islands (Lazio) revealed a good ecological status of coastal water. However, in Zannone Island, the available chemical analysis indicates a higher value than the Carlit index. In

[1]Water Framework Directive 2000 /60 /EC
[2]Ecological Quality Ratio

this work we investigate the role of the superficial *Cystoseira amentacea var.stricta* belts (a brown macroalgae bioindicator of water quality, see Table 1). We analyze the relation between the algae presence and the geomorphological characteristics of Pontine Islands coast through standard logistic regression and autologistic models to account for spatial correlation in order to specically evaluate the predictive capacity of these characteristics. Our report is focused on the Island of Zannone.

2 Materials and Methods

The survey was carried out through a small boat at 3-4 meters of distance from the coastline. The recorded data were obtained noting, by use of a GPS, the discontinuities of the coast concerning algal communities and geomorphological characteristics. Thus the obtained sample units are homogeneous coastal sectors. For each Island the following observed data are given: (1)Population Category (Cod-popol), label of the observed community as described in table 1; (3) Coastal Morphology categorized as BM-metric blocks, FA-high cliff, FB-low cliff and SP-beach; (4) Sensitivity to pollution level (Value), (see table 1);(5) Length of the homogenous coastal sector in meters (Length). All the information has been transferred in ArcGis software. Exploratory data analysis has been carried out for all the variables and the presence of *Cystoseira* was coded as a binary variable (0/1). The association between the latter and the observed covariates (Slope and Morphology) has been explored trought he χ^2 test. The variable Slope, in this step, has been categorized as 0-30, 45-60, 75-90. These data have been aligned to a Digital Elevation Model layer with resolution 20 meters, superimposing a grid of 326 cells to Zannone coastline. This new dataset allows us to investigate the association between the algae presence and a more detailed evaluation of the Island morphology. In the combined dataset the presence of each category of *Cystoseira* coded as a binary variable, the slope, elevation and aspect of the coast are available. The coast slope and aspect have been coded into 3 and 8 categories respectively for exploratory purposes, while in models estimation slope is taken back to its original expression.
Standard logistic regression models (Agresti, 2002) are estimated on the combined dataset to evaluate the predictive capacity of morphological GIS information for the algae presence with and without discrimating by population category. As standard logistic regression does not account for spatial autocorrelation, that seems a natural feature of this type of data, results are then compared with the predictive capacity of autologistic models (Besag, 1974) estimated through pseudo-likelihood (Besag 1975, Huang and Ogata 2002). To predict algae presence each grid cell with model-estimated probability of presence larger then 0.5 is set to 1.

Category	Description	Sen. level
Cystoseira brachycarpa/crinita /elegans	Community dominated by *Cystoseira brachycarpa /crinita /elegans*	20
Cystoseira sheltered	Community dominated by *Cystoseira foeniculacea /barbata /humilis / spinosa*	20
Cystoseira amentacea /mediterranea 5	Continuous belt of *Cystoseira mediterranea /stricta*	20
Cystoseira amentacea /mediterranea 4	Almost continuous belt of *Cystoseira mediterranea /stricta*	19
Cystoseira amentacea /mediterranea 3	Abundant patches of dense stands of *Cystoseira mediterranea /stricta*	15
Cystoseira amentacea /mediterranea 2	Abundant scattered plants of *Cystoseira mediterranea /stricta*	12
Cystoseira compressa	Community dominated by *Cystoseira compressa*	12
Cystoseira amentacea /mediterranea 1	Rare scattered plants of *Cystoseira mediterranea /stricta*	10

Table 1: Summarized description and sensitivity levels of the community categories related to *Cystoseira* as reported in the methodological contribution published by ISPRA (Mangialajo et al. 2008).

3 Results

The exploratory data analysis of the observed data reported a not significant association between the presence of *Cystoseira* and the coast Morphology at the available detail, while significant relation is found with slope. When considering a more detailed representation of the Island, as given by the DEM layer, significant association are found for all *Cystoseira* communities and the morphological variables slope and aspect, categories with higher sensitivity values showing stronger association. Logistic for the algae presence without discriminating by category return high significance of slope and aspect. Through this model 23.31% of the predicted grid cells where misclassified (with 37 wrong 1's and 39 wrong 0's over 326 grid cells). For categories with sensitivity level 19 and 15 we obtain similar results with a misclassification error of 15.95% with 12 wrong occurrences and 40 wrong zeros, showing a tendency to underestimate the number of presences. For less sensitive communities the significance of slope and aspect is reduced and the logistic model produces a 26.38% of missclassified cells, with a stronger tendency to underestimate the algae presence.
within the pseudolikelihood estimation approach, the autologistic model corresponds to a logistic regression model in which the number of occurrences in each cell's neighborhood (SV) is a regressor. In this study a simple first order neighborhood is adopted. For all community categories the SV shades the relations of the algae presence with slope and aspect that become not significant, however the predictive capability of such models is considerably enhanced, misclassification errors drop considerably (for the general presence of *Cystoseira* 8.59%, for highly sensitive com-

munities 2.76% and for less sensitive communities 7.36%), all models tend to slightly underestimate the presence. A poisson regression (Cameron, Trivedi, 1998)has been fitted to relate SV and slope and aspect. Results show the strong relation between these variables explaining why the presence of SV in the model hides the dependence of the algae presence on the other two.

4 Concluding remarks

All analysis confirm the relevance of morphological variables in determining the *Cystoseira* communities presence in the Zannone island with stronger influence on more sensitive ones. Accounting for spatial correlation allows a considerably more precise prediction. Future work will deal with the remaining islands of the Pontine archipelago. It is of interest to investigate other models under a Bayesian estimation approach.

References

Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley

Ballesteros E., Torras X., Pinedo S., Garcia M., Mangialajos L., De Torres M. (2007) A new methodology based on littoral community cartography dominated by macroalgae for the implementation of European Water Framework Directive, *Marine Pollution Bullettin*, 55, 172-180.

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems, *Journal of the Royal Statistical Society, Series B*, 23, 192-236.

Besag, J. (1975) Statistical Analysis of Non-Lattice Data. *The Statistician*, 24, 179-195

Cameron A.C., Trivedi P.K. (1998). *Regression analysis of count data*, Cambridge University Press

Huang, F. , Ogata, Y. (2002). Generalized Pseudo-Likelihood Estimates for Markov Random Fields on Lattice, *Annals of the Institute of Statistical Mathematics*, 54, 1-18.

Mangialajo L., Sartoni G., Giovanardi F., (2008). Quaderno metodologico sull'elemento biologico macroalghe e sul calcolo dello stato ecologico secondo la metodologia CARLIT. ISPRA, Istituto Superiore per la Protezione e la Ricerca Ambientale, 105

Pinedo S., Garcia M., Satta M.P., De Torres M., Ballesteros E. (2007) Rocky shore communities as indicators of water quality: A case study in Northwestern Mediterranean *Marine Pollution Bullettin*, 55, 126-135.

Statistical issues in the assessment of urban sprawl indices [1]

Daniela Cocchi, Linda Altieri
Statistical Sciences Dept., University of Bologna, daniela.cocchi@unibo.it; linda.altieri@studio.unibo.it

Marian Scott, Massimo Ventrucci
School of Mathematics and Statistics, University of Glasgow, marian.scott@glasgow.ac.uk, massimo.ventrucci@glasgow.ac.uk

Giovanna Pezzi
Experimental Evolutionistic Biology Dept., University of Bologna, giovanna.pezzi2@unibo.it

Abstract: Urban sprawl is a hotly debated issue, even if a universally agreed definition does not exist. Its evaluation on spatial data is very important, but the properties of commonly used landscape and sprawl indices have to be assessed, and their performance on raster maps at different pixel resolutions checked, in order to better understand the uncertainty and reliability of results.

Keywords: urban sprawl indices, land cover, pixel resolution, raster aggregation rules, spatial dependence indices.

1. Introduction

Urban sprawl is an important issue for biologists, urban specialists, planners and statisticians, and also for official statistics, both in developed and new developing countries. A universally accepted, well established definition of urban sprawl does not exist, but one of its fundamental properties is to capture uncontrolled and inefficient urban dispersion, accompanied by low building density. Urban sprawl usually occurs when urban planning is not well managed; among its consequences are high average transport costs, soil sealing, pollution (Bhatta *et al*., 2010). Three main types of urban sprawl are currently under study: the monocentric form (one core city surrounded by sprawled suburbs), the polycentric form (more than one core city) and the decentralised pattern (no city centre).
Various measurement methods have been proposed in recent years (see a review in Bhatta *et al*., 2010); some of them are absolute (based on the choice of a sprawl threshold for a selected index), other relative (comparison-based). A very popular sprawl index is Shannon's entropy, but the literature advises that a set of complementary indices to integrate information is created to give a more precise idea of this complex phenomenon. Each index is calculated with reference to a certain spatial extent and a certain spatial data resolution, and measures can be compared over space/time.

[1] Work supported by the project PRIN 2008: New developments in sampling theory and practice, Project number 2008CEFF37, Sector: Economics and Statistics, awarded by the Italian Government.

Statistics can address the sprawl issue in many ways, especially by evaluating the most common recent sprawl indices, assessing their properties, uncertainty and behaviour on raster datasets. Our aim is to identify a suitable set of sprawl indices with good properties and the ability to distinguish among the three sprawl forms; additional information comes from the study of indices at different aggregation levels, following the two most commonly used aggregation methods: the majority and the random rule (He *et al*., 2002).
In our study we have used official EEA land cover data, from the CORINE Land Cover programme (http://eea.europa.eu). They are collected from nearly all EU countries and consist of vector data; the data are then rasterised to 100x100 and 250x250m pixel resolution; a binary raster dataset is also derived, which divides the land into urbanised and non-urbanised zones.

2. Motivation of simulation and empirical studies

Starting from the same elementary data, indices of urban sprawl can assume different values according to the level (*i.e.* pixel dimension) and the aggregation method. We started with a simulation study, necessary for assessing the non linearity of the problem under study, then we used the real dataset mentioned above to detect sprawl occurrence. Both studies were run on raster binary data.
We chose a small set of spatial and landscape indices (*i.e.* we do not exploit information on population, transport, pollution …) and assessed, by simulation, their statistical properties. Each index has a different function: they indicate the existence of sprawl (Shannon's Entropy and Contagion's Index, the last being a measure of clustering-dispersion), the proportion of the territory involved (Simpson's Evenness) and the kind of sprawl (Moran's I, a measure of spatial dependence, because we believe we should find hardly any spatial correlation among pixels in sprawled areas); the interesting ability of Moran's I to identify the type of sprawl has been hypothesized by Tsai (2005) and verified and confirmed by our simulation study. Shannon's Entropy, is defined as $H = -\sum_{i=1}^{S} p_i \ln(p_i)$, where p_i is the proportion of pixels of class i and S is the total number of classes (2 in the case of binary data). It varies between 0 (no sprawl) and $ln(S)$ (maximum sprawl); the usual threshold for sprawl is $ln(S)/2$ (Bhatta *et al*., 2010).
The proportion error has been computed to check the reliability of pixel aggregation in terms of similarity to the original image. We have aggregated both simulated and real datasets to three levels following both rules, to compare the two methods' performance and see how much error in our indices' results they cause.
Our simulation study has reproduced the three sprawl types in various scenarios, generated by an underlying autologistic model (following Hughes *et al.*, 2010) plus Gibbs sampling method. The classic autologistic model is defined as

$$P(Z_i = 1 \mid Z_{-i}, \theta) = p_i = \frac{\exp\left\{ X_i \beta + \eta \sum_{j \in N(i)} Z_j^* \right\}}{1 + \exp\left\{ X_i \beta + \eta \sum_{j \in N(i)} Z_j^* \right\}}$$

where Z_i is the i-th pixel's response, Z_{-i} are all other responses in the grid and the vector θ includes the spatial and attraction parameters; the covariates X_i are the spatial coordinates, weighted through the spatial β parameter, and the autocovariates Z_j are the neighbours' values, weighted through the attraction parameter η. According to this model, the probability of finding urbanisation in the i-th cell depends only on its neighbours' responses (the relationship is controlled by η) and by the pixel location (through the value of β). To create the core city area, we fixed high values for both parameters, while sprawled areas had negative values for η. The neighbourhood extent $N(i)$ has to be fixed in advance, and we chose the 4 nearest neighbours system (Bivand *et al.*, 2008).
In our simulations, we firstly varied spatial and attraction parameters in the model, and, as an alternative, we imposed a kernel structure from the core city area to the periphery, *i.e.* the pixels' responses and the proportion of urbanised cells depend negatively on the Euclidean distance from the city centre; this second, more realistic hypothesis has led to better and more coherent results. For each scenario (9 as a whole), we produced 1000 replications and aggregated them to two coarser levels with both rules (54 scenarios in total). We have then computed the above indices on all replications and resolutions. The same computations have been done on both Emilia Romagna and the city of Bologna (Figure 1) areas (selected from CORINE data) where the original datasets have been aggregated to 500x500,1000x1000 and 5000x5000m pixel sizes.

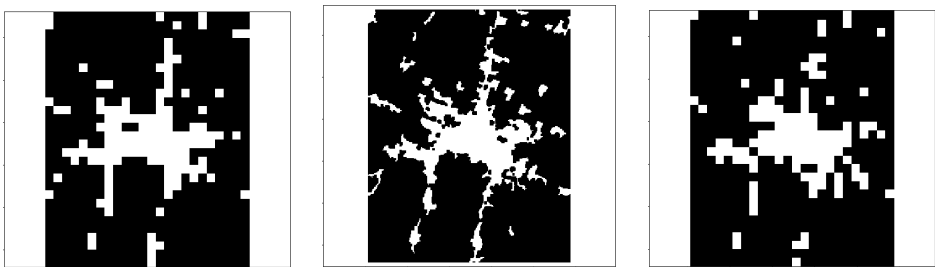


Figure 1. Bologna datasets in the original resolution (100x100m, central panel) and aggregated to 1x1km with the majority method (left panel) and random method (right panel) .

3. Results and comments

Results evaluate indices' stability along aggregation levels and methods, to respectively assess the bias induced by a loss of pixel resolution, and/or using a different aggregation method. The majority rule is a deterministic aggregation method, while the random rule basically draws a simple random sample for each aggregation, starting from the finer resolution data. It appears to be very reliable in the dichotomous case, because the probability of an aggregated pixel falling into one binary class is proportional to the percentage of original pixels in the population of finer elements. The majority method tends to cluster and over-represents the pixels with higher frequency: it is not suitable for detecting dispersion in the data, because it will tend to underestimate it. This has been noted, *e.g.*, in the simulation results for Shannon's Index: after two aggregation steps with the majority rule, the Index did not show occurrence of sprawl, completely contradicting the results from the original data. In conclusion the random aggregation rule is good for measuring sprawl, and leads, in general, to very stable results, *i.e.* more similar to the original, even if its variability always has to be considered.

The variability in indices' measures (in simulation analysis, measured with standard errors and ranges) is higher the coarser the resolution, irrespective of the aggregation method: this suggests it is better to work on the finest resolution possible, even if results are stable over aggregation. The proportion error, which is a classification error, also increases when the resolution becomes coarser, but this tendency is stronger with the majority rule than the random rule. Simpson and Shannon's measures lead to analogous results because they are both based on urbanised pixels' proportions; since no information on pixels' spatial distribution is used, we suspect that they are not the best in identifying sprawl. They are stable when aggregating with the random rule, and, with real data, they identify sprawl in Bologna but not in Emilia Romagna, which suggests that these indices are not reliable on such a wide spatial extent: sprawl is a metropolitan, not a regional, problem. The contagion measure, which is a modified entropy measure containing some information on pixels' neighbourhood, is consistent with Shannon's I, remains stable with the random rule and states that there is sprawl in Bologna. Moran's I is the only index which is able to distinguish (in our kernel simulation study) among the three sprawl types, as shown in Figure 2; on real data it detects occurrence of monocentric sprawl in Bologna, as supported by its map visualization (Figure 1).

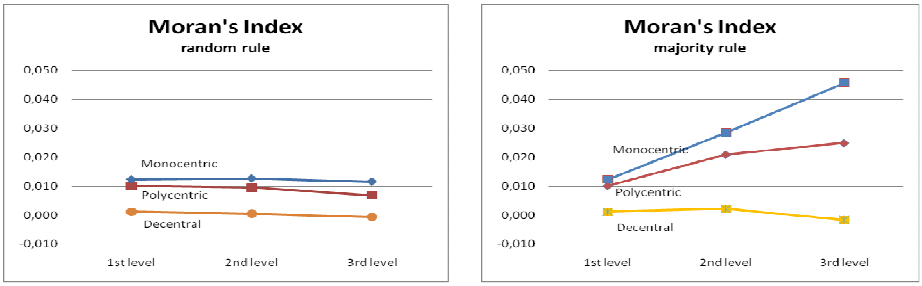


Figure 2. Kernel simulation study; Moran's I at various scenarios and aggregation levels, with both aggregation rules.

In conclusion, the chosen set of indices is suitable for measuring urban sprawl and for identifying the type of dispersion; a further step will be the construction of a unique, composite indicator to identify and quantify such spatial sprawl. As CORINE original data are in vector form, indices such as Simpson, Shannon and Moran's (for binary data) Index should be also computed on vector data to check consistency among results.

References

Bhatta, B., Saraswati, S. and Bandyopadhyay, D. (2010). Urban sprawl measurement from remote sensing data. *Applied Geography*, 30, 731–740.

Bivand, R.B., Pebesma, E.J. and Gómez-Rubio, V. (2008). Applied spatial data analysis with R, UseR! Series, Springer.

He, H.S., Ventura, S.J. and Mladenoff, D.J. (2002). Effects of spatial aggregation approaches on classified satellite imagery. *International Journal of Geographical Information Science*, 16(1), 93-109.

Hughes, J., Haran, M. and Caragea, P. (2010). Autologistic models for binary data on a lattice. Environmetrics.

Tsai, Y.H. (2005). Quantifying Urban Form: Compactness versus 'Sprawl'. *Urban Studies*, 42(1), 141–161.

Using spatial statistics tools on remote-sensing data to identify fire regime linked with savanna vegetation degradation

Jacquin Anne, Chéret Véronique
Université de Toulouse, INPT - Ecole d'Ingénieurs de Purpan, UMR 1201 DYNAFOR, 75 voie du TOEC, BP57611, F-31076 Toulouse Cedex 3, France, anne.jacquin@purpan.fr

Goulard Michel
INRA, UMR 1201 DYNAFOR, Chemin de Borde-Rouge, Auzeville, F-31326 Castanet-Tolosan, France

Sheeren David
Université de Toulouse, INPT – ENSAT, UMR 1201 DYNAFOR, BP32607, F-31326 Castanet-Tolosan, France

Abstract: Fire is acknowledged to be a factor for explaining the disturbance of vegetation dynamics interacting with other environmental factors. Depending on the fire regime, the amount of herbaceous biomass changes but depends on local conditions. We want to clarify the importance and the role of fire on the dynamics of savanna vegetation. The study area is the Marovoay watershed located on the north-west coast of Madagascar. In this watershed, burning herbaceous cover is the main practice in the extensive grazing system. The image dataset is composed of two indicators related to vegetation activity changes and one indicator about fire regime that results from a combination of fire frequency and seasonality. All indicators were measured between 2000 and 2007 using a remote sensing MODIS time series. In this work, we implemented two approaches of spatial analysis. The first one is based on a per-pixel non-spatial GLM model and analyzes the spatial structure of the residuals. In the second approach, a spatial GLM model is directly computed. We built stratifications of the study area according to the spatial variations of the relationship established between vegetation activity changes and fire regime. The use of spatial statistical tools produces parsimonious models which we found to be consistent with expert knowledge.

Keywords: fire regime, spatial statistics, GLM model, vegetation dynamics, remote sensing

A methodology for assessing the spatial distribution of static wildfire risk over wide areas: the case studies of Liguria and Sardinia (Italy)[1]

Antonella Bodini, Erika Entrade
Institute of Applied Mathematics and Information Technology
(CNR-IMATI, Milano), antonella.bodini@mi.imati.cnr.it

Q. Antonio Cossu, Simona Canu
Environmental Protection Agency of Sardinia (ARPAS)

Paolo Fiorucci, Francesco Gaetani
CIMA Research Foundation

Ulderica Paroli
Regione Liguria, Civil Protection and Emergency Department

Abstract: In Mediterranean areas, some studies suggest universal increases in fire frequency due to climatic warming. However, some authors point out that the universality of these results is questionable. In this study, we try to go beyond the simple analysis of statistical data related with the number of fires and the total burned area, which can be misleading in the context of climate change. The fire perimeters have been used to inquire spatialized climate indexes and the vegetation cover. A statistical analysis of climate indexes has been conducted and a certain number of Type of Homogeneous Areas (THA) defined by introducing information on vegetation cover. The comparison of THA and climatic indexes allowed the definition of an index of risk. Maps of this index highlight risky areas in Liguria and Sardinia (Italy).

Keywords: climate change, climate indexes, static wildfire risk, vegetation cover.

1. Introduction

In Mediterranean area, some studies in the later ’90 (Piñol *et al*. 1998) predicted a continue increase of the number of days of very high fire risk, and more frequent catastrophic wildfires. Some studies, in the same period, suggested universal increases in fire frequency with climatic warming (Overpeck *et al*. 1990). However, Flannigan *et al*. (2000) point out that “the universality of these results is questionable because an individual fire is a result of the complex set of interactions that include ignition agents, fuel conditions, topography and weather including temperature, relative humidity, wind velocity and the amount and frequency of precipitation. Increasing temperature alone does not necessarily guarantee greater fire disturbance.”.

[1] This work has been supported by the project PROTERINA-C: A system for the forecast and the prevention of the impact of the variability of the climatic conditions on the risk for the natural and urbanized environment. Funded by the EU (2009-2011), “Obiettivo 3 Italia-Francia Marittimo” program.

In this study, we try to go beyond the simple analysis of statistical data related with the number of fires and the total burned area which can be misleading in the context of climate change. The availability of a long data series of fire perimeters combined with a detailed knowledge of topography and land cover allow to understand which are the main features involved in forest fire occurrences and their behaviour. In addition, the analysis of climate indexes allows to understand the role of climate on fire regime, both in terms of direct effects on fire behaviour and the effect on vegetation cover.

2. Materials and Methods

Study areas. Liguria (Italy) is a region of 5400 km^2 lying on the northwest coast of the Tyrrhenian Sea. For this Mediterranean region, wildfires are recurrent phenomena both in summer and winter: an average of 365 wildfires of size > 0.01 km^2 burns an area of 55 km^2 per year.
Sardinia (Italy) is the second-largest island in the Mediterranean Sea. Wildfires represent a severe threaten to life and goods during summer. On average, between May and October more than 2500 fires burn more than 310 km^2 of shrubland, grassland and forests per year.
Data. As far as fire perimeters are concerned, the data set used in Liguria references the period from 1997 to 2009 and reports 7390 wildland fires that overall burnt 510 km^2 of forests and shrubland. The dataset used in Sardinia references the period from 2006 to 2008 and reports more than 4850 fires that overall burnt 480 km^2. The available regional vegetation cover maps are different in the two regions, then preventing a homogeneous classification. Daily rainfall and temperature data referring to the period 1951-2008 have been analyzed using time series a) with more than 30 years of complete records for trend analysis, and b) with more than 20 years of complete records in the standard period 1971-2000, for climate analysis.
Method. Climate indexes have been analyzed at the seasonal and annual temporal scale. In particular, the maximum number of consecutive dry days (CDD) and the heat wave duration index (HWDI) suggested by Frich *et al*. (2002) for monitoring change in climatic extremes world-wide have been considered. Interpolated maps of the normal values have been obtained by either kriging or multiple regression. For each climate index, a finite number of classes has been defined on the basis of a preliminary analysis of the fire perimeters. These classes have been compared to land cover classes to derive the possible Types of Homogeneous Areas (THA). The index:

$$\mathbf{H_i} = \frac{1}{\Delta T} \frac{\sum_{k=1}^{N_i} \mathbf{B_{ik}}}{\mathbf{A_i^{tot}}}$$

has been defined to measure the risk associated to the *i*-th THA, where ΔT is the length of data series (years), N_i the number of fires occurred in THA *i*-th, $\mathbf{B}_{ik}$ is the burned area in THA *i*-th by fire *k*-th, $\mathbf{A}_i^{tot}$ is the total cover area by THA *i*-th.

3. Results

In general, trend analysis does not show clear patterns of climate change at the annual time-scale. As far as extreme events in Sardinia are concerned, CDD shows decreasing trend in the Central-South part, where consecutive dry days periods are longer. Similar results are not obtained in Liguria, where, on the contrary, 3 stations show increasing CDD.
In Liguria, the actual number of THA is 906 in winter and 865 in summer, while in Sardinia the number of actual THA is 395. Figure 1 shows the obtained maps of index **H** and highlights the areas at the highest risk (red).

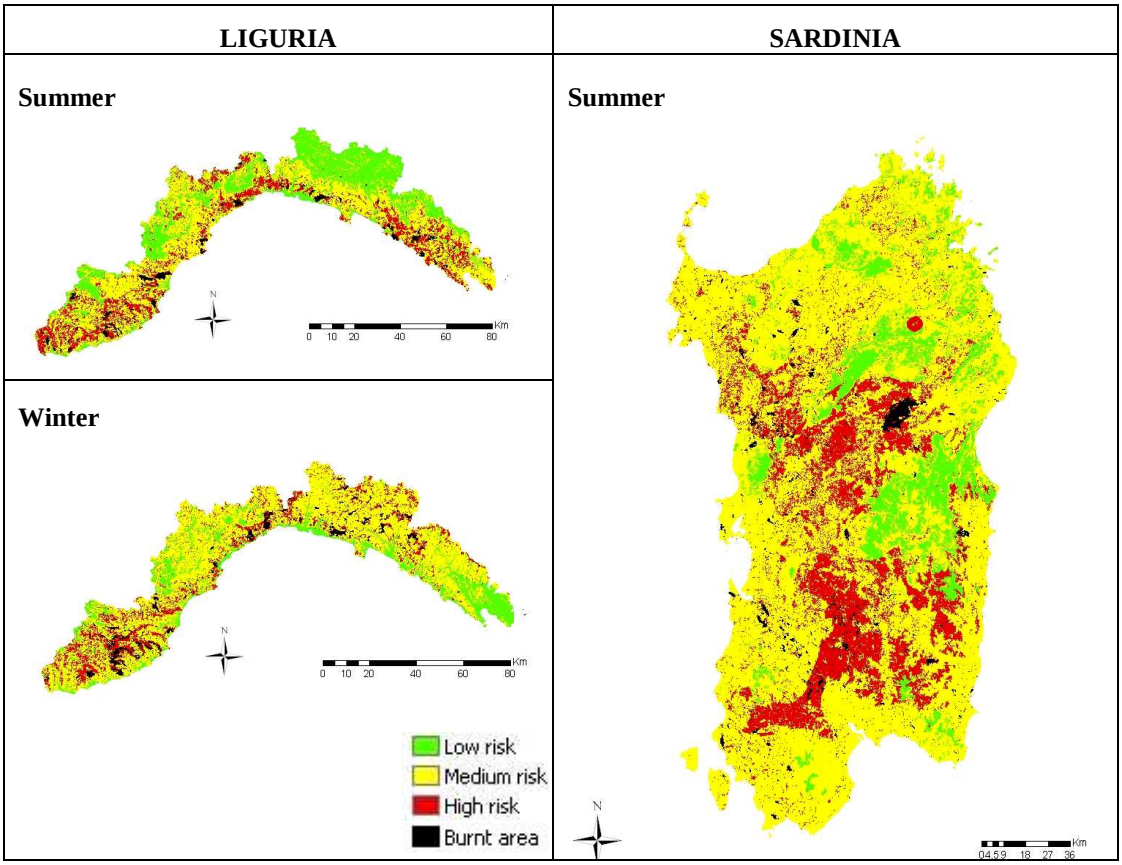


Figure 1: Risk maps in Liguria (left) and Sardinia (right). In black, the burned areas are shown.

4. Concluding remarks

In this work, an index of static fire risk has been introduced that takes into account both the climate and the vegetation cover. Two different case studies are presented, Regione Liguria and Regione Sardegna (Italy). Both regions are in the center of the Mediterranean and are characterized by a high number of fires and burned area. However, the two regions have very different fire regimes. Sardinia is affected by the

fire phenomenon only in summer whilst Liguria is affected by fires also in winter, with higher number of fires and larger burned area. In addition, the two region are very different in vegetation cover.
Concerning Liguria, the proposed methodology is able to put in evidence the different seasonal fire regimes, and provides useful information about regional fire risk management. Shrublands, in Liguria, represent the first stage of the succession dynamics in abandoned agriculture areas and it is the most flammable kind of vegetation mainly involved in large fires both in summer and in winter season. Conifer plantations near the coastline mainly constituted by *Pinus Pinaster* heavily degraded by Matsucossus represent the most high intensity and frequent fires in summer which provide the major risk for the Wildland-Urban Interface. On the contrary, in Sardinia, shrubs represent less that 10% of the vegetation cover contained in the higher risk areas and conifers are not present at all. Here, the vegetation cover characterizing the higher risk areas is mainly composed by *Quercus Suber* and mixed forests.
Concerning the climate influence on fire risk, in summer season the highest risk areas are characterized in both regions by high air temperature. Only in Sardinia the HWDI seems to play a key role. The influence of rainfall regime on fire risk in the summer season (Sardinia is affected by fires only in summer) puts in evidence, in both regions, that the higher risk areas are characterized by a significant amount of total precipitation and by a significant number of rainy days. However, the same areas are characterized by a significant number of cumulative dry days especially in Sardinia. This result shows that fire ignition is mainly favorite by the presence of annual herbaceous species which accumulate biomass in the wet season and represent almost a completely dry fuel in the summer season.
The obtained results are certainly satisfying, however suggest further improvements. The role of HWDI has to be discussed, and a seasonal definition introduced. Indeed, the relevant role of this index in Sardinia only can be due to the higher temperatures of the island. Further insight into the role of HWDI and CDD has been obtained by comparing the dates of fires and the corresponding indexes values. We obtained a strong relationship between higher HWDI values and long dry periods in spring. This result open the doors to further development of the analysis.

References

Flannigan M.D., Stocks B.J., Wotton B.M. (2000) Forest fires and climate change, *Science of the Total Environment*, 262, 221-230

Frich P., Alexander L.V., Della-Marta P., Gleason B., Haylock M., Klein Tank A., Peterson T. (2002) Global changes in climatic extremes during the 2nd half of the 20th century, *Climate Research*, 19, 193-212

Piñol, J., Terradas J., Lloret F. (1998) Climate Warming, Wildfire Hazard, and Wildfire Occurrence in Coastal Eastern Spain, *Climate Change*, 38, 345-357.

Overpeck J.T., Rind D., Goldberg R. (1990) Climate-induced changes in forest disturbance and vegetation, *Nature* 343, 51 - 53

A new procedure for fitting a multivariate space-time linear coregionalization model [1]

Sandra De Iaco
Dip.to di Scienze Economiche e Matematico-Statistiche, Facolta' di Economia, Universita' del Salento, Italy, sandra.deiaco@unisalento.it

Monica Palma
Dip.to di Scienze Economiche e Matematico-Statistiche, Facolta' di Economia, Universita' del Salento, Italy.

Donato Posa
Dip.to di Scienze Economiche e Matematico-Statistiche, Facolta' di Economia, Universita' del Salento, Italy.

Abstract: New classes of cross-covariance functions have been recently proposed, nevertheless the linear coregionalization model (LCM) is still of interest and widely applied. In this paper, a new fitting procedure of the space-time LCM (ST-LCM) using the generalized product-sum model is proposed. This procedure is based on the well known algorithm of matrix simultaneous diagonalization, applied on the sample matrix variograms computed for multiple spatial-temporal lags.

Keywords: spatial-temporal correlation, product-sum variogram model, linear coreginalization model.

1 Introduction

The LCM, firstly introduced by Matheron in 1982, is still one of the most utilized models for multivariate spatial and spatial-temporal data analysis (Zhang, 2007; Babak and Deutsch, 2009; Emery, 2010). However, in the space-time context several theoretical and practical aspects must be considered, such as the fitting process. In geostatistics, there is a wide literature concerning the LCM fitting stage (Goulard and Voltz, 1989; Lark and Papritz, 2003). In this paper, a new fitting procedure of the ST-LCM using the generalized product-sum variogram model is proposed. It is shown that the simultaneous diagonalization of the sample matrix variograms is useful to identify the basic components of the coregionalization model.

[1]Supported by Fondazione Cassa di Risparmio di Puglia.

2 Multivariate space-time random field

Given a second-order stationary vector-valued space-time random function (*STRF*) $\{\mathbf{Z}(\mathbf{s},t),\ (\mathbf{s},t) \in D \times T \subseteq \mathbb{R}^{d+1}\}$, with $\mathbf{Z}(\mathbf{s},t) = [Z_1(\mathbf{s},t),\ldots,Z_p(\mathbf{s},t)]^T$, $p \geq 2$, where $\mathbf{s} = (s_1, s_2, \ldots, s_d) \in D$ (generally, $d \leq 3$), denotes the spatial coordinates and $t \in T$ is the temporal coordinate, the cross-variogram of two space-time random functions $Z(\mathbf{s},t)$ and $Z(\mathbf{s}',t')$ exists and depends on the space-time separation vector $\mathbf{h} = (\mathbf{h}_s, h_t)$, with $\mathbf{h}_s = (\mathbf{s}-\mathbf{s}')$ and $h_t = (t-t')$. As in the spatial context, a second-order stationary multivariate *STRF* can be modelled as a *ST-LCM*. Hence, the variogram matrix can be written as

$$\boldsymbol{\Gamma}(\mathbf{h}) = \boldsymbol{\Gamma}(\mathbf{h}_s, h_t) = \sum_{l=1}^{L} \mathbf{B}_l\ g_l(\mathbf{h}_s, h_t), \tag{1}$$

where $\mathbf{B}_l = [b^l_{\alpha\beta}],\ l = 1,\ldots,L,\ \alpha,\beta = 1,\ldots,p$, are positive definite $(p \times p)$ matrices, commonly known as *coregionalization matrices*, while $g_l(\mathbf{h}_s, h_t),\ l = 1,\ldots,L$, are basic space-time variograms associated with the L scales of variability.
In De Iaco et al. (2003, 2005), each space-time basic variogram is modelled as a generalized product-sum model (De Iaco et al., 2001):

$$g_l(\mathbf{h}_s, h_t) = \gamma_l(\mathbf{h}_s, 0) + \gamma_l(\mathbf{0}, h_t) - k_l\ \gamma_l(\mathbf{h}_s, 0)\ \gamma_l(\mathbf{0}, h_t)\ , \quad l = 1,\ldots,L, \tag{2}$$

where $\gamma_l(\mathbf{h}_s, 0)$ and $\gamma_l(\mathbf{0}, h_t)$ are the spatial and temporal marginal variogram models, respectively, while parameters $k_l, l = 1,\ldots,L$, are given by:

$$k_l = \frac{sill[\gamma_l(\mathbf{h}_s,0)] + sill[\gamma_l(\mathbf{0},h_t)] - sill[g_l(\mathbf{h}_s,h_t)]}{sill[\gamma_l(\mathbf{h}_s,0)] \cdot sill[\gamma_l(\mathbf{0},h_t)]}, \qquad l = 1,\ldots,L. \tag{3}$$

By substituting (2) in (1), the *ST-LCM* based on the generalized product-sum variogram models is determined by two marginal *LCM*'s:

$$\boldsymbol{\Gamma}(\mathbf{h}_s, 0) = \sum_{l=1}^{L} \mathbf{B}_l\ \gamma_l(\mathbf{h}_s, 0), \qquad \boldsymbol{\Gamma}(\mathbf{0}, h_t) = \sum_{l=1}^{L} \mathbf{B}_l\ \gamma_l(\mathbf{0}, h_t). \tag{4}$$

Note that other space-time variogram models (Gneiting, 2002; Ma, 2002; Stein, 2005; Porcu et al., 2008) can be used to describe the basic components of the *ST-LCM*. However, the flexibility of the product-sum variogram, in estimating and modeling the spatial-temporal variability, is often convenient (De Iaco et al. 2003, 2005).

3 Fitting a *ST-LCM*

After a brief review of the usual fitting process of the *ST-LCM* using the generalized product-sum model, the new, more flexible, fitting procedure is discussed.

The usual fitting procedure

In De Iaco et al. (2003) the process of fitting a *ST-LCM* using a generalized product-sum variogram model, was developed as follows.

1. Compute the empirical marginal direct variograms, in space and in time, for all the p variables under study and then fit nested variogram models. At this step, the diagonal elements of each matrix $\mathbf{B}_l,\ l = 1, \ldots, L$, are determined as well as the marginal basic structures $\gamma_l(\mathbf{h}_s, 0)$ and $\gamma_l(\mathbf{0}, h_t)$, $l = 1, \ldots, L$.
2. Determine the marginal cross-variograms and the off-diagonal elements of the matrices (4), ensuring that each matrix $\mathbf{B}_l$ is positive definite.
3. In order to complete the modeling of $g_l(\mathbf{h}_s, h_t)$, $l = 1, \ldots, L$, the k_l parameters must be determined. Hence, the space-time variogram surfaces are computed and fitted to product-sum nested models.

Using this procedure, different practical problems have to be faced: a) the identification of the $b_{ij}^l, i, j = 1, \ldots, p$, elements of the matrices $\mathbf{B}_l,\ l = 1, \ldots, L$, since for a fixed l, these coefficients must be the same for the marginal space and time variograms; b) the estimation of parameters k_l, with $l = 1, \ldots, L$.

The new fitting procedure

Given the multivariate space-time data set concerning the p variables (with $p \geq 2$) and the $p(p+1)/2$ spatio-temporal direct and cross-variograms, computed for a selection of H spatial-temporal lags, the new fitting algorithm goes on running 4 sub-procedures sequentially, as follows.

Sub-procedure I: identify the basic structures.

A simultaneous diagonalization technique is applied on the set of H square, symmetric and real-valued matrices $\hat{\mathbf{\Gamma}}(\mathbf{h}_s, h_t)_k, k = 1, \ldots, H$, of sample direct and cross-variograms, in order to find a $(p \times p)$ orthogonal matrix which diagonalizes or "nearly" diagonalizes these matrices. At this step, the l-th empirical basic spatial-temporal component are detected by extracting the l-th diagonal element from all the diagonal matrices.

Sub-procedure II: fit the basic structures.

Given the space-time surfaces of the basic components, the spatial and temporal ranges of the basic surfaces are determined so that the scales of space-time variability are identified. The number L ($L \leq p$) of scales depends on the number of different spatial and temporal ranges the basic components exhibit. Successively, the product-sum model $g_l(\mathbf{h}_s, h_t)$ in (2) is fitted to each empirical basic component, with $l = 1, \ldots, L$. Hence marginal variogram models, $\gamma_l(\mathbf{h}_s, 0)$ and $\gamma_l(\mathbf{0}, h_t)$ are fitted to the empirical basic marginals.

Sub-procedure III: compute the coregionalization matrices.

Given the direct and cross-variograms surfaces of the variables under study, estimated in step I, the global sill values at the L scales of spatial-temporal variability are detected. Successively, the elements $b_{\alpha\beta}^l$ of matrices $\mathbf{B}_l, l = 1, \ldots, L$, are determined by dividing the contributions of the direct and cross-variogram surfaces at the l-th scale of variability by $sill[g_l(\mathbf{h}_s, h_t)]$.

Sub-procedure IV: check the admissibility of the model.
Given the coregionalization matrices $\mathbf{B}_l, l = 1, \ldots, L$, the admissibility of the *ST-LCM* is checked. If the matrix $\mathbf{B}_l$, with $l = 1, \ldots, L$, presents some negative eigenvalues, they are replaced by zeros, such that the new coregionalization matrix $\mathbf{B}_l^+$, at the l-th scale of variability, is positive definite.

References

Babak, O., Deutsch, C.V. (2009) An intrinsic model of coregionalization that solves variance inflation in collocated cokriging, *Comput. & Geosc.* 35, 3, 603-614.

De Iaco, S., Myers, D.E., Posa, D. (2001) Space-time analysis using a general product-sum model, *Statist. and Probab. Lett.*, 52, 1, 21-28.

De Iaco, S., Myers, D.E., Posa, D. (2003) The linear coregionalization model and the product-sum space-time variogram, *Mathematical Geolology*, 35, 1, 25-38.

De Iaco, S., Palma, M., Posa, D. (2005) Modeling and prediction of multivariate space-time random fields, *Comput. Statist. and Data Anal.* 48, 525-547.

Emery, X. (2010) Iterative algorithms for fitting a linear model of coregionalization, *Comput. & Geosc.*, 36, 9, 836-846.

Gneiting, T. (2002) Nonseparable, stationary covariance functions for space-time data, *J. of the Am. Statist. Assoc.*, 97, 458, 590-600.

Goulard, M., Voltz, M. (1989) Linear coregionalization model: tools for estimaton and choice of cross-variogram matrix, *Math. Geol.*, 24, 269-286.

Lark, R.M., Papritz, A. (2003) Fitting a linear model of coregionalization for soil properties using simulated annealing, *Geoderma*, 115, 245-260.

Ma, C. (2002) Spatio-temporal covariance functions generated by mixtures, *Math. Geol.*, 34, 965-975.

Porcu, E., Mateu, J., Saura, F. (2008) New classes of covariance and spectral density functions for spatio-temporal modelling, *Stoch. Environ. Res. and Risk Assess.*, 22, Supplement 1, 65-79.

Stein, M. (2005) Space-time covariance functions, *J. of the Am. Statist. Assoc.*, 100, 310-321.

Zhang, H. (2007) Maximum-likelihood estimation for multivariate spatial linear coregionalization models, *Environmetrics* 18, 125-139.

Bayesian hierarchical models : An analysis of Portugal road accident data [1]

Conceição Ribeiro
ISE-UALG/CEAUL, cribeiro@ualg.pt

Antónia Amaral Turkman
DEIO-FCUL/CEAUL

João Lourenço Cardoso
LNEC

Abstract: In this work Bayesian hierarchical models are applied to road accident data at a county level, in Portugal, from 2000 to 2007. The objective of the study is to build model-based risk maps for road accidents at county level and to perform an analysis of association between road accidents and potential risk factors, through the inclusion of ecological covariates in the model.

Keywords: Bayesian models; Small Area; Road Safety.

1 Introduction

Investigation into risk factors relating to road safety and transport plays an important role in road accident analysis and prevention. Heterogeneity in road accidents can be related to a range of factors, in particular at a small area level. Bayesian hierarchical models allow the incorporation of spatial and temporal effects through prior information and enable ecological analyses of associations between road accidents and potential risk factors over aggregated areas. One objective is to explore and to examine potential associations between road accidents and regional characteristics. Here we only consider three covariates namely the road length, population county size and county area.

2 Materials and Methods

Consider $Y_{ij} \sim Poisson(E_{ij}\theta_{ij})$, where, for each county i and each year j, Y_{ij} is the observerd number of fatal and severe injury crashes, θ_{ij} is the relative risk, E_{ij} is the expected number of fatal and severe injury crashes, for a constant incidence rate

[1] This work is partially sponsored by national funds through FCT - Fundação para a Ciência e a Tecnologia under the project PEst-OE/MAT/UI0006/2011 and by SFRH/PROTEC/49226/2008 PhD grant.

across all 278 counties of Portugal and all 8 years,

$$E_{ij} = N_{ij}\bar{r} = N_{ij}\frac{\sum_i \sum_j Y_{ij}}{\sum_i \sum_j N_{ij}} \tag{1}$$

with N_{ij} the number of vehicles insured in county i, in year j.

Assume a spatio-temporal model for the relative risk, namely

$$\log(\theta_{ij}) = b_0 + bx_{ij} + u_i + v_i + (\gamma + \delta_i)t_j \tag{2}$$

where b_0, bx_{ij} are the fixed effects,with b_0 the intercept, x_{ij} a vector of covariates, b a vector of fixed effect parameters; u_i are random effects acounting for spatial heterogeneity and v_i are random effects accounting for unstructured heterogeneity. We also assume that u_i and v_i are mutually independent with priors $v_i \overset{iid}{\sim} Normal(0, \sigma_v^2)$ and the $u_i \sim CAR$, respectively; γt_j is a linear trend term in time t_j, δ_i is an interaction random effect between space and time, with prior $\delta_i \sim CAR$. We also assumed the following diffuse priors for the hyperparameters: b_0, b, σ_u^2, and σ_v^2 mutually independent with $b_0, b, \gamma \sim Normal(0, 1000)$, $(\sigma_u^2)^{-1}, (\sigma_v^2)^{-1}, (\sigma_\delta^2)^{-1} \sim Gamma(0.5, 0.0005)$.

The models were applied to road accident data in 278 counties of Portugal, from 2000 to 2007 and were implemented using WinBUGS, (Spiegelhalter et al.,1999) and its add-on program GeoBUGS, (Thomas et al., 2004), and using R-INLA, (Rue and Martino, 2009).The covariates used were geographical area-A, in Km^2, population size-P, in number of inhabitants, road lenght-L, in meters, by county and year.

3 Results

Eight spatial-temporal models are implemented. Model 1 without covariates -1-. Models 2 to 4, incorporate one covariate, model 2 includes the road length, 2-(L), model 3 includes the population size, 3-(P), and model 4 includes the area, 4-(A). Models 5 to 7 incorporate two covariates, model 5 includes road length and population size, 5-(L+P), model 6 includes road length and area, 6-(L+A), and model 7 includes population size and area, 7-(P+A). Finally, model 8 incorporates the three covariates, 8-(L+P+A).

Results obtained using INLA and WINBUGS are very similar, with the advantage of INLA taking much less time to run. Model choice is done using DIC (see table 1); accordingly model 2, which has a smaller value for DIC, is chosen to produce maps to display the posterior expected relative risk. Figure 1 display, on the left side, the observed average of fatal and severe injury crashes along the years under study and on the right side the expected relative risks obtained using model 2.

ST Models		1	2-(L)	3-(P)	4-(A)
DIC	INLA	12799.9	12796.1	12800.2	12803.6
	WB	12799.8	12796.0	12799.6	12803.2
ST Models		5-(L+P)	6-(L+A)	7-(P+A)	8-(L+P+A)
DIC	INLA	12796.7	12796.7	12803.8	12797.0
	WB	12796.7	12797.0	12803.6	12797.3

Table 1: DIC values

		mean	sd	0.025q	0.975q
Fixed effects:	b_0	0.03	0.05	0.02	0.04
	b_L	1.6e-06	4.9e-07	6.5e-07	2.6e-06
	γ	-0.088	0.003	-0.095	-0.082
Variance of random effects:	σ_u^2	0.34	0.07	0.18	0.55
	σ_v^2	0.07	0.02	0.04	0.13
	σ_δ^2	0.007	0.001	0.004	0.011

Table 2: Summary statistics for the parameters in model 2

Figure 1: Portugal: Average of fatal and severe injury crashes and posterior expected relative risks for model 2 in 2000

Apparently the model is not able to capture well the risk of accident on the northwest coast line, indicating that other covariates should be considered to be included in the analysis. Summary statistics for the fixed effects and the variance

of the random effects for model 2, in table 2, show that road length can be a factor associated with higher risk of accident. The time trend effect being negative may be an indicator that the number of fatal and severe injury crashes decreased over the study period. The analysis of the variance of the random effects shows that the variability of the relative risk is attributed more to spatial-structured effects than to the uncorrelated heterogeneity or to the space-time interaction.

4 Concluding remarks

This is a preliminary analysis, as we are well aware that there are potential risk factors that are not accounted for in this study. These include socioeconomic factors such as age cohorts, sex cohorts, levels of poverty and employment; transportation-related factors such as road type, road curvature, traffic flow, traffic speed, violation of traffic rules, number of vehicle-kilometers traveled, and environmental factors such as total precipitation, number of rainy days per year, land use, size of rural and urban areas.

References

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., and Songini, M. (1995) Bayesian analysis of space-time variation in disease risk, *Statistics in Medicine*, 14, 2433-2443.

Ghosh, M., Natarajan, K., Waller, L.A., and Kim, D. (1999) Hierarchical bayes GLMs for the analysis of spatial data: An application to disease mapping, *Journal of Statistical Planning and Inference*, 75(2), 305- 318.

Miaou, S.-P., Song, J.J., and Mallick, B.K. (2003) Roadway traffic crash mapping: A space-time modeling approach, *Journal of Transportation and Statistics*, 6(1), 33-57.

Rue, H. and Martino, S. (2009) Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations, *Journal of the Royal Statistical Society B*, 71(2), 319-392.

Spiegelhalter, D.J., Thomas, A., and Best, N.G. (1999) WinBUGS Version 1.2 User Manual, Technical report, MRC Biostatistics Unit.

Thomas, A., Best, N., Lunn, D., Arnold, R., and Spiegelhalter, D. (2004). GeoBUGS User Manual Version 1.2, Technical report, Department of Epidemiology and Public Health of Imperial College at St Mary's Hospital London.

Electrical resistivity measurements for spatial soil moisture variability estimation

Giuseppe Calamita, Raffaele Luongo
IMAA-CNR, Contrada Santa Loja Zona Industriale - 85055 Tito Scalo, (Pz), Italy
DIFA, Università della Basilicata, c.da Macchia Romana – 85100, Potenza, Italy
calamita@imaa.cnr.it

Angela Perrone, Vincenzo Lapenna, Sabatino Piscitelli, Salvatore Straface
IMAA-CNR, Contrada Santa Loja Zona Industriale - 85055 Tito Scalo, (Pz), Italy

Abstract: This experimental work empirically compares the results obtained in soil moisture spatial estimation performed with different interpolation techniques. Three algorithms were compared: Inverse Distance Weight (IDW), Ordinary Kriging (OK) and Co-Kriging (CoK). The data used were obtained through an *in-situ* sampling in a test site located in central Italy. The calibration and the validation data set contain respectively 40 and 133 point measurements of TDR soil moisture. The covariate used is a data set of 533 electrical resistivity (conductivity) point measurements. In terms of prediction accuracy results show no great differences between the performance of the IDW and the OK methods. Quite more accurate results were obtained incorporating the secondary variable information in the CoK algorithm.

Keywords: TDR, electrical resistivity, geostatistic, ordinary kriging, co-kriging.

1. Introduction

Being a key variable in many natural processes acting at different spatial/temporal scales, there is a great interest in the observation, estimation and interpretation of soil moisture (SM) patterns. Traditionally the in-situ measurements have been performed by using the thermo-gravimetric method and, most recently, Time Domain Reflectomery (TDR) and neutron probes. These methods can be very precise and accurate but they are invasive and carry information representative only for small areas and volumes. Emerging electrical resistivity (ER) method has been applied in a growing number of surveys. This technique is relatively less invasive, cost effective and gives information of a larger volume of soil. Our interest is on the investigation of SM spatial variability using different interpolation algorithms, both deterministic (Inverse Distance Weight, IDW) and geostatistical. Moreover, we would like to compare punctual soil moisture predictions obtained by Ordinary Kriging (OK) method, applied on the poor sampled SM variable, with those obtained through the Co-Kriging algorithm (CoK) incorporating the more dense information of the secondary variable (ER).

2. Materials and Methods

The study area is a 200m x 60m test site located in the Umbria region (central Italy). Simultaneous measurements of SM [% vol/vol] and ER [Ohm*m] were acquired on the

nodes of a 5m sampling step regular grid. A mobile TDR probe with 15 cm wave-guide length (MiniTrase, Soil Moisture Equipment Corporation) was used for the SM measurements. A geo-resistivimeter Syscal Junior (IRIS Instrument) coupled with a Schlumberger 4-electrode device was used for the ER measurements at ~20cm of pseudo-depth.
The focus of this work was to compare different interpolation techniques in order to verify the advantages of using auxiliary variables for soil moisture patterns estimation. A strong under sampling of the SM variable was performed (Fig1) as often is in real cases. Through a completely random sampling, 10 calibration data sets (40 points each) and a validation data set (133 points) for each calibration data set were sub-sampled. We present here the results concerning one of the ten calibration data sets (Fig1).

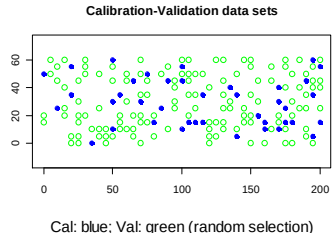


Figure 1: Relative positions of points used for the validation and one of the calibration data set.

Three different interpolation algorithms, IDW, OK and CoK (Govaerts, 1997; Hengl, 2007), were applied to obtain SM spatial predictions. All the interpolations and variogram models were performed on log transformed variables and then data were back-transformed for the validation of the results.
To fit a linear model of co-regionalization under the constrain of having positive definite partial sill matrices, we chose to use electrical conductivity, (EC= 1/ER), values [mS/m] because the experimental variogram looked more similar to the SM one.
Two validation steps were applied: first the difference between predicted and measured values (validation data set) was computed; then, a regression between predicted and measured values was conducted and residuals compared in terms of: mean prediction error (MPE), median, standard deviation (RMSE), mean absolute error (MAE), minimum and maximum value. Moreover, the Pearson correlation coefficients between predicted and measured values of SM for each interpolation algorithm were estimated.

3. Results

The summary statistics (Tab.1) show that SM variability (sd and CV) is broader for the validation than for the calibration dataset. The central values are slightly lower for the validation than for the calibration set. The log transformation of data was applied in order to obtain quasi-normal pdfs.

variable	unit	count	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	CV	skewness
sm: calibration	% vol/vol	40	21,30	27,48	29,30	30,40	32,78	44,10	5,17	0,17	0,78
sm: validation	% vol/vol	133	17,60	26,40	28,90	29,83	34,10	45,90	5,26	0,18	0,31
electrical conductivity	mS/m	533	3,70	13,09	16,90	17,96	21,69	39,31	6,76	0,38	0,55

Table 1: Summary statistics of the different data sets.

The distributions of the EC, the SM validation and calibration sets are showed in Fig.2 along with a scatter plot of the logEC-logSM relation for the calibration sample.

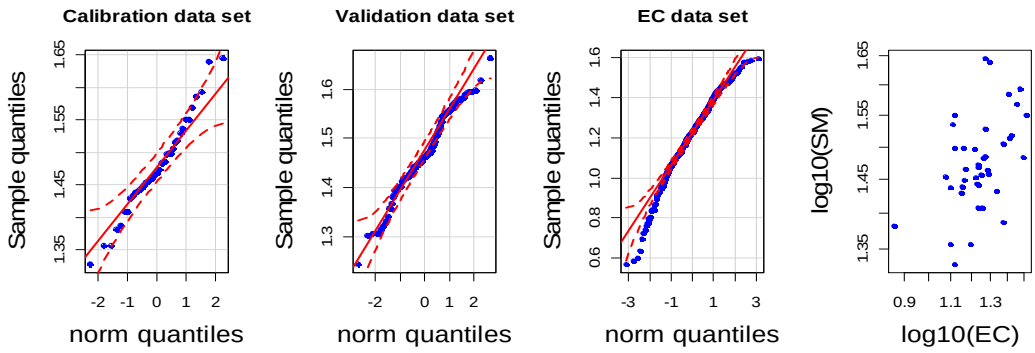


Figure 2: data set distributions and log(SM) vs log(EC) correlation.

After having modeled the SM variogram, the OK algorithm for spatial interpolation was applied and the results compared with those obtained through the IDW. The visual comparison of the two maps (not showed) highlighted no strong differences. The general large scale patterns are similarly reproduced. The OK map seemed to be smoother and with less artifacts. The statistics comparing the experimental errors were only slightly better for the OK than for the IDW algorithm (Tab.2).

(measured - predicted)	MPE	Median	RMSE	MAE	I quart	III quart	Min	Max	r
IDW	0,59	0,52	4,45	3,48	-2,10	4,45	-9,99	12,16	0,54
OK	0,56	0,64	4,43	3,47	-2,20	3,21	-9,86	10,70	0,55
CoK	0,38	-0,09	3,82	2,97	-1,74	2,66	-9,86	9,55	0,70

Table2: residual summaries between measured and predicted values.

Once the EC spatial structure and its covariance with SM were modeled (co-regionalization modeling) (Tab.3), a CoK map was produced. The map well reproduced the larger SM patterns and clearly showed better defined SM smaller scale details (Fig.3).
Moreover, the correlation between CoK predicted and validation data was sharply higher than that for OK and IDW (0.70 vs ~0.55) and the error statistics showed a better performance of the CoK in modeling the SM values (Tab.2).

	model	psill	range (m)
log(SM)	Nug	6,45E-04	
	Exp	3,57E-03	13
log(EC)	Nug	3,00E-03	
	Exp	2,98E-01	13
log(SM)log(EC)	Nug	3,17E-03	
	Exp	1,06E-02	13

Table 3: Model parameters of cross- and semi- variograms for the log-transformed variables

4. Concluding remarks

The comparison between different interpolation algorithms in modeling the spatial SM patterns was shown. The first comparison was done between OK and IDW algorithms, both accounting for the SM sampled data (40 points). These results were compared with that of the CoK algorithm that allowed us to incorporate the information of a more densely sampled covariate (530 data). The validation of the three interpolation methods applied was assessed using an independent validation data set (133 points).
The results of the validation procedure showed no marked differences between IDW and OK performances, with the latter being just slightly more accurate than the former. On the other hand, the comparison in term of SM estimation revealed the major accuracy of the CoK algorithm respect to the IDW and OK algorithms. This result confirm the relevance that secondary variables can have in spatial modeling. Further analysis of various issues need to be explored: how sample size and scheme affect the spatial predictions of the OK and CoK algorithms? And, how interpolation algorithms perform compared to simulations in modeling the spatial variability of the SM?

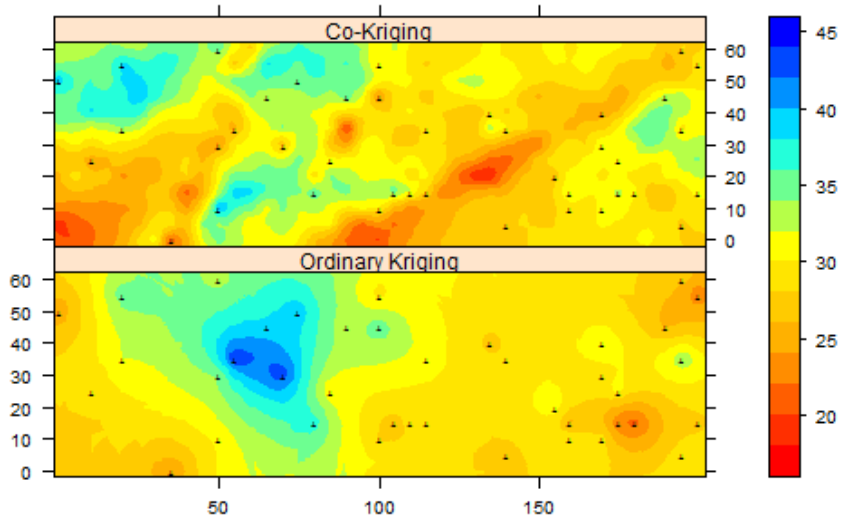


Figure 3: CoK map (top) compared with OK map (bottom) of the SM spatial pattern in the test site. Symbols on the map indicate the calibration sampling sites.

References

Boureanne H., King D., Couturier A., Nicollaud B., Mary B., Richard G., (2007). Uncertainty assessment of soil water content spatial patterns using geostatistical simulations: an empirical comparison of a simulation accounting for single attribute and a simulation accounting for secondary information. *Ecological Modelling*, 205.

Goovaerts P, (1997) *Geostatistics for Natural Resources Evaluation*, Oxford University Press, USA

Hengl T., (2007) *A practical guide to geostatistical mapping*, EUR 22904 EN Scientific and Technical Research series, Office for Official Publications of the European Communities, Luxemburg.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences, 30: 683-691.

Geostatistics and GIS: tools for environmental risk assessment[1]

Sabrina Maggio, Claudia Cappello, Daniela Pellegrino
Dip.to di Scienze Economiche e Matematico-Statistiche, Facolta' di Economia,
University of Salento, Italy, sabrina.maggio@unisalento.it

Abstract: The environmental risk analysis involves the observation of complex phenomena. Different kinds of information, such as environmental, socio-economic, political and institutional data, are usually collected. In this paper, spatial-temporal geostatistical analysis is combined with the use of a Geographic Information System (GIS): the integration between geostatistical tools and GIS enables the identification of alternative scenarios and possible strategies for the environmental risk management. A case study on environmental data measured in the southern part of Apulia region (South of Italy), called Grande Salento, is discussed. Sample data (concentrations of PM_{10}, wind speed, temperature) taken at different air monitoring stations are used for stochastic prediction, through space-time indicator kriging.

Keywords: GIS, Geostatistics, PM_{10} pollution, space-time indicator kriging

1 Introduction

Environmental risk management involves the integrated use of several tools and techniques, including GIS, sample design, Geostatistics and data management. In particular, data management process requires the integration of several data divided into three categories: i) environmental data (land use, land cover, vegetation, geology, meteorology and measures of pollutants concentration); ii) socio-economic data (population and housing census data, community vulnerability data and data on utilities and access); iii) political and institutional data (Chen et al., 2003). Moreover, a spatial-temporal approach is often required for environmental risk assessment; hence, the interaction between space-time modeling of air pollution, adopted by the statistical community in environmental studies (De Iaco et. al, 2001; Kolovos et al., 2004; Spadavecchia and Williams, 2009, among others), and urban environment representation (traffic network, location of industrial facilities, emission sources and topographic conditions), easily managed in a GIS, is necessary. The aim of this paper is to combine the use of space-time geostatistical techniques and the GIS potential. A case study on an environmental data set, involving both atmospheric variables and air pollutant concentrations, measured in November 2009 at monitoring stations located in Grande Salento (Lecce, Brindisi and Taranto districts in the Apulia Region) is discussed. In particular, air pollution due to PM_{10} (Particulate matter) concentrations and atmospherical variables, such as wind speed and

[1]Supported by Fondazione Cassa di Risparmio di Puglia.

temperature in the same region, are considered. Exploratory Spatial Data Analysis for a deep understanding of the analyzed phenomenon is performed using the Geostatistical Analyst Tool of ArcGis. Structural analysis for space-time variogram estimating and modeling and space-time prediction, based on kriging, is computed by using modified *Gslib* routines. A $3D$ representation for the space-time evolution of the conditional probability associated with PM_{10} is produced by using *ArcScene* (an extension of ArcGis). The overlay between the probability map and relevant urban spatial data is shown for Brindisi Municipality.

2 Empirical framework and methods

The study of the evolution of PM_{10} is very important for the effects that this pollutant has on human health. Many studies have shown that exposure to PM_{10} increases the risk of mortality both in long and short term. According to National Laws concerning the human health protection, PM_{10} hourly average concentrations cannot be greater than 50 μg/m^3 for more than 35 times per year. During the month under study, the PM_{10} hourly values exceeded the threshold 80 times, especially on the 13rd, 14th, 23rd and 24th of November. In the present case study, the following steps have been considered: (1) defining the space-time indicator variables according to appropriate thresholds, computed from the observed data; (2) modeling space-time indicator variogram of the variables by using the generalized product-sum variogram model; (3) using space-time indicator kriging, over the area of interest and during the period 1-6 December 2009, in order to obtain: a) the joint probability that PM_{10} concentrations exceed fixed thresholds and the atmopheric variables take values not greater than the corresponding monthly means, b) the joint probability that the atmopheric variables take values not greater than the corresponding monthly means; (4) computation and $3D$ representation of the probability that PM_{10} concentrations exceed the fixed thresholds, conditioned to adverse atmospheric conditions (i.e. wind speed and temperature which are lower than the corresponding monthly mean values). In Geostatistics, observations are modelled as a partial realization of a spatio-temporal random function Z, which is decomposed into a sum of a trend component and a stochastic residual component. In the following case study, the formalism of a spatio-temporal indicator random function ($STIRF$),

$$I(\mathbf{u}, z) = \begin{cases} 1 & \text{in case of } Z \text{ not greater (or not less) than the threshold } z, \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathbf{u} = (\mathbf{s}, t) \in D \times T, z \in \mathbb{R}$ ($D \subseteq \mathbb{R}^2$ and $T \subseteq \mathbb{R}_+$), is considered. Spatio-temporal dependence of a $STIRF$ is characterized by the indicator variogram of I: $2\gamma_{_{ST}}(\mathbf{h}) = \mathrm{Var}[Y(\mathbf{s} + \mathbf{h}_s, t + h_t) - Y(\mathbf{s}, t)]$, which depends solely on the lag vector $\mathbf{h} = (\mathbf{h}_s, h_t)$, $(\mathbf{s}, \mathbf{s} + \mathbf{h}_s) \in D^2$ and $(t, t + h_t) \in T^2$. The fitted model for $\gamma_{_{ST}}$ must satisfy an admissibility condition in order to be valid. Hence, the following generalized product-sum model (De Iaco et al. 2001) has been fitted to the empirical

indicator space-time variograms:

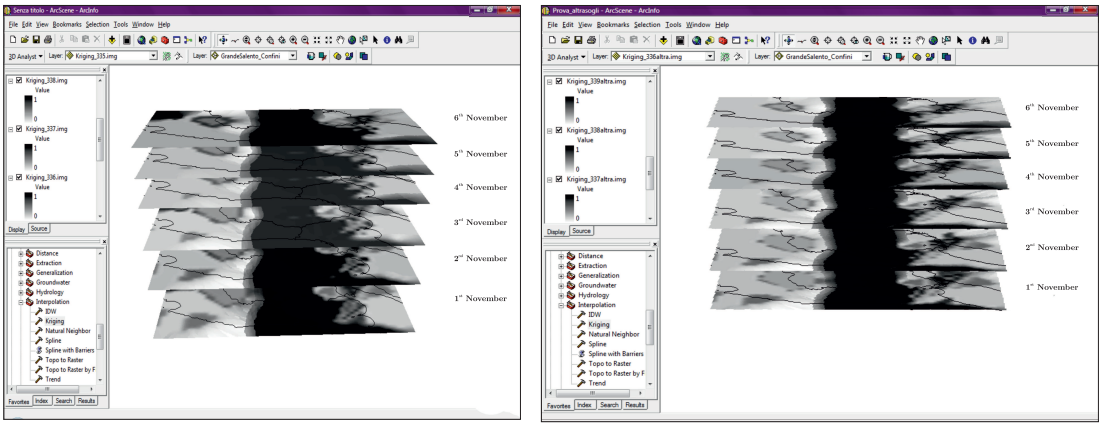
$$\gamma_{ST}(\mathbf{h}_s, h_t) = \gamma_{ST}(\mathbf{h}_s, 0) + \gamma_{ST}(\mathbf{0}, h_t) - k\gamma_{ST}(\mathbf{h}_s, 0)\gamma_{ST}(\mathbf{0}, h_t), \quad (1)$$

where $\gamma_{ST}(\mathbf{h}_s, 0)$ and $\gamma_{ST}(\mathbf{0}, h_t)$ are valid spatial and temporal bounded marginal variograms and $k \in]0, 1/max\{sill\gamma_{ST}(\mathbf{h}_s, 0), sill\gamma_{ST}(\mathbf{0}, h_t)\}]$. Basic theoretical results can be found in De Iaco et al. (2001), moreover recently it was shown that strict conditional negative definiteness of both marginals is a necessary as well as a sufficient condition for the product-sum (1) to be strictly conditionally negative definite (De Iaco et al., 2011).

3 Case study

In this analysis, the $STIRF$s associated with the spatial-temporal distributions of PM_{10}, as well as of temperature and wind speed, have been examined in the Grande Salento region during November 2009. The data set consists of daily averages of 3 variables, PM_{10}, temperature and wind speed, measured in November 2009 at 28 monitoring stations located in the Grande Salento.
After computing descriptive statistics, spatial-temporal indicator kriging using the generalized product-sum variogram model has been applied in order to predict, over the area of interest and for the period 1-6 December 2009, the probability that PM_{10} concentrations exceed the fixed limits, in the presence of adverse atmospherical conditions to the pollutant dispersion, i.e. temperature (T) and wind speed (WS), which are lower than the corresponding monthly mean values (12.54 °C and 2.11 *meters/second*, respectively). In this case study, the thresholds for the PM_{10} have been fixed equal to the 75th and 80th percentiles of samples data (37.804 and 40.57 $\mu g/m^3$, respectively), which can be considered critical with respect to the law limit. Hence, 3 indicator random fields have been defined: $I_1(\mathbf{u}; 37.804, 12.53, 2.11) = 1$, if $PM_{10} \geq 37.804$, $T \leq 12.53$, $WS \leq 2.11$, 0 otherwise, $I_2(\mathbf{u}; 40.57, 12.53, 2.11) = 1$, if $PM_{10} \geq 40.57$, $T \leq 12.53$, $WS \leq 2.11$, 0 otherwise, $I_3(\mathbf{u}; 12.53, 2.11) = 1$, if $T \leq 12.53$, $WS \leq 2.11$, 0 otherwise, with $\mathbf{u} \in D$.
Indicator sample space-time variograms for the indicator variables under study and their models have been determined first. The fitted space-time variogram model for the random fields I_1 is characterized by: $\gamma_{ST}(\mathbf{h}_s, 0) = 0.066\,[1 - \exp(-3\,\mathbf{h}_s/15000)]$, $\gamma_{ST}(\mathbf{0}, h_t) = 0.185\,[1 - \exp(-3\,h_t/6)]$, $k = 3.767$ and global sill equal to 0.205;
for I_2: $\gamma_{ST}(\mathbf{h}_s, 0) = 0.059\,[1-\exp(-3\,\mathbf{h}_s/15000)]$, $\gamma_{ST}(\mathbf{0}, h_t) = 0.169\,[1-\exp(-3\,h_t/6)]$, $k = 4.112$ and global sill equal to 0.187;
for I_3: $\gamma_{ST}(\mathbf{h}_s, 0) = 0.094\,[1-\exp(-3\,\mathbf{h}_s/20000)]$, $\gamma_{ST}(\mathbf{0}, h_t) = 0.235\,[1-\exp(-3\,h_t/6)]$, $k = 3.712$ and global sill equal to 0.247. Probability maps have been predicted over the area of interest for the period 1-6 December 2009. In particular, the indicator kriging has been used to estimate the joint probability that PM_{10} concentrations exceed fixed thresholds and the atmopheric variables take values not greater than the corresponding monthly means first, and secondly the joint probability that the atmopheric variables take values not greater than the corresponding monthly means.

Then, the probabilities that PM_{10} values do not exceed the fixed thresholds, conditioned to adverse atmospheric conditions, over the area of interest and during the period 1-6 December 2009, have been computed. From the obtained results it is evident that, in Brindisi and Lecce districts, the probability that PM_{10} daily concentrations exceed the fixed thresholds, under adverse atmospherical conditions, decreases from the 1st to the 6th of December 2009 and along the NorthWest-SouthEast direction (Fig. 1). Finally, the Brindisi Municipality is considered in



a) b)

Figure 1: conditional probability maps of PM_{10} concentrations, for the thresholds: a) 37.804 $\mu g/m^3$ (75th percentile), b) 40.57 $\mu g/m^3$ (80th percentile), during the period 1-6 December 2009.

detail. In this area, the concentrations of PM_{10} is compared with land use and traffic network. The probability that PM_{10} concentrations do not exceed the fixed threshold is higher in the city center; on the other hand it is much more likelihood that the PM_{10} concentrations exceed the limit in the SouthWest hinterland.

References

Chen K., Blong R., Jacobson C., (2003) Towards an integrated approach to natural hazard risk assessment using GIS: with reference to bushfires, *Environmental Management*, 31, 546-560.

De Iaco, S., Myers, D.E., Posa, D. (2001) Space-time analysis using a general product-sum model, *Statistics and Probability Letters*, 52, 1, 21-28.

De Iaco, S., Myers, D.E., Posa, D. (2011) On strict positive definiteness of product and product-sum covariance models, *Journal of Statistical Planning and Inference* 141, 1132-1140.

Kolovos A., Christakos G., Hristopulos D.T., Serre M.L., (2004) Methods for generating non-separable covariance models with potential environmental applications, *Advances in water resources*, 27, 815-830.

Spadavecchia L., Williams M., (2009) Can spatio-temporal geostatistical methods improve high resolution regionalisation of meteorological variables?, *Agricultural and Forest Meteorology*, 149, 6-7, 1105-1117.

How to estimate anisotropic attenuation exploiting prior isotropic knowledge

Renata Rotondi
C.N.R. - Istituto di Matematica Applicata e Tecnologie Informatiche, Milano (I),
reni@mi.imati.cnr.it

Gaetano Zonno
Istituto Nazionale di Geofisica e Vulcanologia - sez. Milano-Pavia, (I),
zonno@mi.ingv.it

Abstract: The pattern of the highest intensities in macroseismic fields of volcanic areas is strongly anisotropic because of the linear extension of the fault. In the isotropic approach to the estimation of the probability distribution of the site intensity the analysis starts considering the sites inside circular bins, with fixed width, around the epicentre. To consider the source effect it seems natural to shift epicentre to the rupture length and circular bins to elliptical ones. To exploit prior information on the attenuation trend in Italian seismological and volcanic areas we transform the plane so that an ellipse becomes a circle with diameter equal to its minor axis, and then estimate the probability distribution of the site intensity applying the method proposed in Zonno et al. (2009) to the transformed data points.

Keywords: Bayesian estimation, binomial-beta model, macroseismic intensity, elliptical isoseismal

1 Introduction

The problem of the macroseismic intensity attenuation and its variation as a function of the distance from the source is a key factor in the seismic hazard assessment. The standard procedure consists in applying linear regressions which express the site intensity I_s mainly as a function of the epicentre-site distance and of the epicentral intensity I_0. Avoiding the use of any empirical attenuation relationship, Rotondi and Zonno (2004) proposed a probabilistic model for I_s calibrated by exploiting information from zones that are assumed homogeneous from the attenuation point of view. This debatable assumption is dropped in Zonno et al. (2009) and replaced by a hierarchical agglomerative clustering method - implemented by the agnes function of the cluster library of R software - by which a set of Italian well-documented macroseismic fields was decomposed in three classes that are homogeneous from the viewpoint of attenuation. For each of these classes the distribution of I_s was estimated conditioned on I_0 from VII to XI degree of the Mercalli-Cancani-Sieberg (MCS) scale under the assumption of symmetric decay around the epicentre. Since

the seismic attenuation in volcanic environment is very quick because of the much fractured ground and of the very shallow seismicity activity, the Italian volcanic areas were excluded from Zonno et al. (2009) and examined separately in Rotondi et al. (2009). It turned out that a decay, similar to the one characterizing the class with the quickest attenuation, is recorded at a distance reduced by a coefficient 10 in Etna and Vesuvius-Ischia areas, and 2 in Aeolian Islands and Albani Hills. In this article, to take into account the source effect, that is the asymmetric decay, without losing the gathered knowledge, we apply a plane transformation to the intensity data points so as to go back to the circular case. Then we assign the prior distributions of the Bayesian paradigm on the basis of the previous studies, update the model parameters and finally associate the probability functions estimated in this way with the original points.

2 Binomial-beta model for I_s

We recall briefly the probabilistic model proposed in Rotondi and Zonno (2004): as the decay $\Delta I = I_0 - I_s$ is a discrete variable belonging to the domain $\{0, I_0 - 1\}$, it is reasonable to choose for I_s, at a fixed distance from the epicentre, the binomial distribution $Bin(i_s|I_0, p)$ conditioned on I_0 and p and then restrict the support to be $\{1, I_0\}$ by defining $Pr\{I_s = 1\} = Pr\{I_s \leq 1\}$. Moreover, since the ground shaking may differ even among sites located at the same distance, we consider p as a random variable which follows a Beta distribution $Beta(p; \alpha, \beta)$. To insert our prior knowledge on the problem, we draw L distance bins R_j, $j = 1, 2 \ldots, L$, of fixed width around the epicentre of the earthquake and assume that, in all the sites within each jth distance bin, I_s has the same binomial distribution with parameter p_j, which, in turn, follows the distribution Beta$(\alpha_{j0}, \beta_{j0})$. On the basis of the comparison with the class of earthquakes with the quickest attenuation in Italy, we assign to the prior parameters α_{j0}, β_{j0} the values α_j, β_j of that class (note that in the present case the bins are 10 times narrower). Then, through the macroseismic fields of 17 earthquakes observed on the flanks of Mt. Etna, we update the parameters in the distance bin R_j and estimate p_j through the posterior mean:

$$\hat{p}_j = E(p_j \,|\, \mathcal{D}_j) = \frac{\alpha_{j,0} + \sum_{n=1}^{N_j} i_s^{(n)}}{\alpha_{j,0} + \beta_{j,0} + I_0 \cdot N_j}$$

where N_j is the total number of data points $\mathcal{D}_j$ inside R_j and $i_s^{(n)}$ is the intensity felt at the n-th site. In order to let the p parameter of the binomial distribution for the site intensity I_s vary with continuity, we smooth the estimates $\hat{p}_j$ with the method of least squares, using an inverse power function $g(d) = (\gamma_1/d)^{\gamma_2}$. In this way it is possible to assign the binomial probability of I_s at any d distance from the epicentre and to use the mode of this distribution to forecast the intensity that could be felt at that distance from the epicentre of a future event of intensity I_0.

To judge the predictive power of this model we perfom a retrospective analysis by comparing the observed macroseismic fields with the predicted ones on the basis of three validation criteria: the logarithmic scoring rule, the odds ratio and the absolute discrepancy between observed and estimated intensities at site. This last criterion can be also used to validate the best empirical attenuation relationship $\Delta I = 0.98 \log d + 1.01$ given in the literature for the Etna volcanic area. The smallest values of the criteria indicate the model with the best performance.

2.1 Anisotropic model

When there is evidence of a preferential direction of propagation, it can be reasonable to assign elliptical isoseismal contours. To this end we apply a transformation to the plane so that the ellipse of major axis equal to the rupture length and minor axis equal to 1 km (width of the distance bin) becomes the unit circle. In this way we can apply the same probabilistic model defined in the isotropic case to the so-transformed data points, estimate a new probability distribution of the site intensity, and associate the new estimates with the original locations. An example of this transformation is depicted in Figure 1 and consists of the following steps: rotate the ellipse of semimajor and semiminor axis $a = 2.023$ and $b = 1$ long respectively, and azimuth 2.356 (135°) counterclockwise by 0.785 rad (45°) so as to move the point $P_1(x_1, y_1)$ to the point $P_2(x_2, y_2)$ through the equations $x_2 = \cos(-\psi)\ x_1 - \sin(-\psi)\ y_1$ and $y_2 = \sin(-\psi)\ x_1 + \cos(-\psi)\ y_1$, being ψ the angle between the positive semi-axis and the directrix. Then shrink the major axis bringing $P_2(x_2, y_2)$ to $P_3(x_3, y_3)$ by $x_3 = x_2 \times b/a$ and $y_3 = y_2$; finally rotate the circle clockwise so that the point P_3 goes to the point $P_4(x_4, y_4)$ by the equations $x_4 = \cos(\phi)\ x_3 - \sin(\phi)\ y_3$ and $y_4 = \sin(\phi)\ x_3 + \cos(\phi)\ y_3$, being $\phi = \arctan(y_3/x_3) - \arctan(y_2/x_2) + |\psi|$. Since the asymmetry is more evident for the highest intensities and decreases when moving away from the epicenter, both the axes of the subsequent ellipses are increased by the same quantity, - width of the bin - so that the eccentricity $e_j = \sqrt{1 - (b_j/a_j)^2}$ tends towards 0 for increasing j, $j = 1, 2, \ldots, L$.

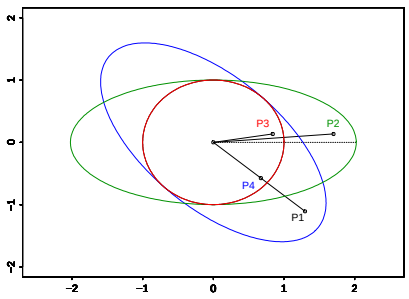


Figure 1: Transformation of the ellipse (2.023,1) into the circle with radius 1, azimuth = 2.356, $\psi = -0.785$ rad.

3 Simulation of damage scenarios

In both isotropic and anisotropic case the method presented allows a complete treatment of the uncertainty; as a matter of fact not only the estimate of the intensity at any site is obtained but also its entire probability distribution from which it is possible to draw additional information, like the probability that the expected intensity exceeds a fixed degree and, viceversa, the intensity that will be felt with a fixed probability threshold. Figure 2 shows an application of the method to the simulation of the damage scenarios that could be generated by an earthquake of intensity IX in the circular (left) and elliptical (right) hypothesis.

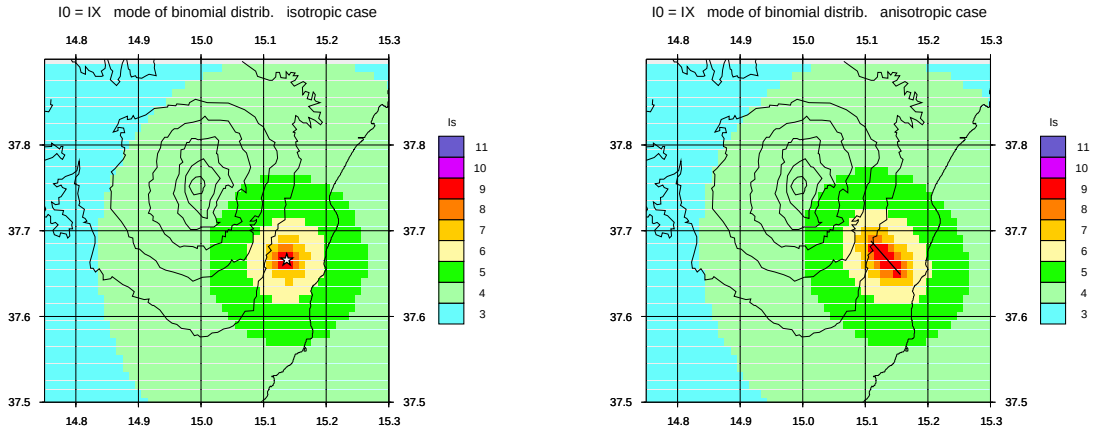


Figure 2: Damage scenarios simulated in isotropic and anisotropic case for $I_0 = \text{IX}$.

Acknowledgements

This work was funded by the Italian Dipartimento della Protezione Civile in the frame of the 2007-2009 Agreement with Istituto Nazionale di Geofisica e Vulcanologia - INGV, project V4: "Hazard connected to the flank dynamics of Etna".

References

Rotondi R. and G. Zonno (2004). Bayesian analysis of a probability distribution for local intensity attenuation, *Annals of Geophysics*, 47, 5, 1521-1540.

Rotondi R., Brambilla C., Zonno G., Azzaro R., D'Amico S., Tuvè T. (2009). Classification of Macroseismic Fields and Forecasting of Damage Scenarios Caused by Earthquakes in Italian Volcanic Districts, *Proceed. ECOBIOSYS 2009 - Classification and Forecasting Models*, Milano, May 15 2009, 4 pp.

Zonno G., Rotondi R., Brambilla C. (2009). Mining Macroseismic Fields to Estimate the Probability Distribution of the Intensity at Site, *BSSA*, **98**, 5, 2876-2892

Natural radioactivity distribution and soil properties: a case study in southern Italy [1]

Ilaria Guagliardi, Nicola Ricca, Maria Grazia Cipriani, Donatella Civitelli, Raffaele Froio, Anna Lia Gabriele, Gabriele Buttafuoco

National Research Council of Italy - Institute for Agricultural and Forest Systems in the Mediterranean (ISAFOM), Via Cavour 4-6, Rende - Cosenza (Italy) ilaria.guagliardi@isafom-cnr.it

Rosanna De Rosa

Department of Earth Sciences, University of Calabria, Ponte Pietro Bucci, 87036 Arcavacata di Rende - Cosenza (Italy)

Abstract: Mapping environmental radioactivity from field gamma-ray spectrometry is a valuable tool for understanding and interpreting pedological control of naturally occurring radioactivity. Soil properties and water content affect the behaviour of natural radioactivity. The main aim of the study were to explore and map the activity of three naturally occurring radionuclides (^{232}Th, ^{238}U, ^{40}K) in an olive orchard and investigate the relationship between some soil properties and the activity of the three radionuclides.

Keywords: natural radioactivity, soil, water content, grain size

1. Introduction

Environmental natural radioactivity in the soil is due to the decay of radionuclides derived from minerals in the earth's crust. Many naturally occurring elements have radioactive isotopes, but only potassium, and the uranium and thorium decay series have radioisotopes producing gamma rays of sufficient energy and intensity to be measured by gamma ray spectroscopy (IAEA, 2003). The radioactive isotope of potassium ^{40}K occurs as a fixed proportion of K in the natural environment and these gamma rays can be used to estimate the total amount of K present. Uranium and thorium occurs naturally as the radioisotopes ^{238}U, ^{235}U and ^{232}Th. Neither ^{238}U nor ^{232}Th emit gamma rays and their concentrations are estimate from their radioactive daughter products and reported as equivalent uranium (eU) and equivalent thorium (eTh).

The mineral composition of the parent material controls the natural radioactivity of soils (Navas et al., 2011) and the processes of weathering, sedimentation, leaching and sorption, and the movement of groundwater may influence activity levels of natural radionuclides (Dowdall and O'Dea, 2002). Soils play a major role in the cycling of radionuclides and their physico-chemical properties influence the mobility and bioavailability of these radionuclides in terrestrial ecosystems (Kabata-Pendias and Pendias, 2001). A fundamental characteristic of the soil, which greatly influences the environmental transport of radioactivity, is the distribution by grain size.

Approximately 95% of the measurable gamma radiation is emitted from the upper 0.5 meters of the profile (Gregory and Horwood, 1961) and the value of gamma spectroscopy lies principally in the amount of radioisotopes of K, U and Th contained in rocks and soil profiles (Dickson and Scott, 1997). Signal attenuation of radioactivity increases by approximately 1% for each 1% increase in volumetric soil water content (Cook et al., 1996).

[1] This research was funded by the Action 2 – Public research laboratory mission oriented. APQ – Scientific Research and technological Innovation in Calabria Region. Laboratory for Food Quality, Safety, and Origin (QUASIORA).

Geostatistical methods provide us a valuable tool to study spatial structure of the activity of radionuclides. They take into account spatial autocorrelation of data to create mathematical models of spatial correlation structures commonly expressed by semivariograms. The interpolation technique of the variable at unsampled locations, known as kriging, provides the 'best', unbiased, linear estimate of a regionalized variable in an unsampled location, where 'best' is defined in a least-square sense (Webster and Oliver, 2007).
The main objectives of the study were: (a) to explore and map the activity of three naturally occurring radionuclides (^{232}Th, ^{238}U, ^{40}K) in an olive orchard, (b) to investigate the relationship between some soil properties and the activity of the three radionuclides.

2. Materials and methods

The experimental area (100 m x 100 m) is an olive orchard located in southern Italy (Calabria). Ground measurements of gamma rays were carried out using the portable gamma-ray spectrometer GRM-260 of the GF Instruments®. Each measurement included the full spectrum of the natural gamma-radiation (counts per 4 minutes) and registered in 256 channels, each of which equal to 12 keV. The counts were then transformed into activity of the corresponding radioactive elements. The conventional approach to the acquisition and processing of gamma-ray data is to monitor four broad spectral regions of interest (ROI) corresponding to potassium-40 (ROI_K), uranium-238 (ROI_U), thorium-232 (ROI_{Th}) and the total count (ROI_{TC}).
The gamma ray measurements were made at 361 points at the nodes of a regular 5 x 5 m grid. Volumetric soil water measurements were made at the same locations with 45-cm long rods of a two-probe Trase System TDR (time domain reflectometry) (Topp and Davis, 1985).
Soil samples were collected at only 100 points at the nodes of a regular 10 x 10 m grid, then they were air-dried, ground, passed through a 2 mm sieve and analysed for particle size fractions by the pipette method.
The gamma ray and soil measurements $z(\mathbf{x}_\alpha)$ at different locations $\mathbf{x}_\alpha$ (x is the location coordinates vector and α the sampling points) were interpreted as a particular realization of a random variable $Z(\mathbf{x}_\alpha)$ and analysed using the theory of random functions (Chilès and Delfiner, 1999; Webster and Oliver, 2007; Wackernagel, 2003, among others). As quantitative measure of spatial correlation of the observations $z(\mathbf{x}_\alpha)$, was used the variogram $\gamma(\mathbf{h})$ which is a two-point statistics used to quantify the variability between two random variables separated by a lag vector **h**. Multi-Gaussian ordinary kriging (Verly, 1983) was used to predict and map the gamma rays and soil particle size fractions values at unsampled locations. It allows spatial prediction of soil properties regardless of the shape of the sample histogram. The multi-Gaussian approach is based on a multiGaussian model and requires a prior Gaussian transformation of the initial attribute $\{Z(\mathbf{x}),\ \mathbf{x} \in R^2\}$ into a Gaussian-shaped variable $\{Y(\mathbf{x}),\ \mathbf{x} \in R^2\}$ with zero mean and unit variance. Such a procedure is known as Gaussian anamorphosis (Wackernagel, 2003).

3. Results and conclusions

In opposition to what was expected (Bihari and Dezső, 2008, among many others), no significant correlation was found between soil particle size and gamma ray measurements.
Experimental variograms (Fig. 1) were computed along four directions of azimuth (0, 45, 90, and 135 in degrees clockwise from N-S axis) for gamma-ray data, particle size fractions and soil water content. Then, a variogram model was fitted to the experimental values of semivariance. Figure 1 shows the variograms only for the spectral regions of interest (ROI) corresponding to potassium-40 (ROI_K) and the total count (ROI_{TC}), the soil water content and the soil silt content.

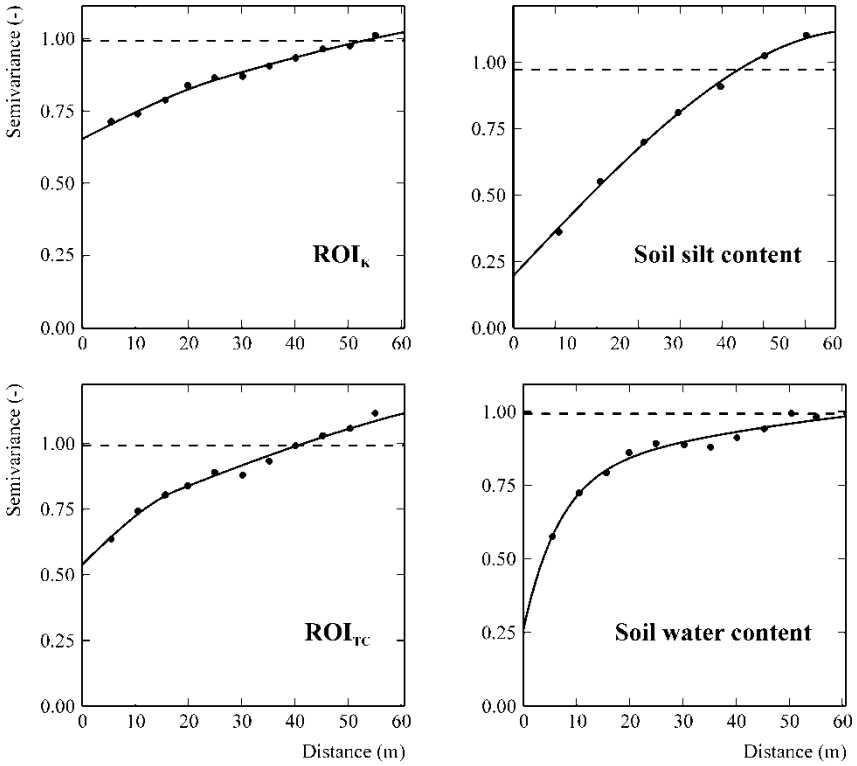


Figure 1: Variograms for the spectral regions of interest (ROI) corresponding to potassium-40 (ROI_K) and the total count (ROI_{TC}), the soil water content and the soil silt content. The experimental values are the filled points and the solid lines are of the model of variograms. The dashed lines are the experimental variances.

For ^{40}K (ROI_K) a nested variogram model was used including three basic structures (Fig. 1): (1) a nugget effect of 0.6529; (2) a spherical model (Webster and Oliver, 2007) with a range of 26 m and a sill of 0.0624; (3) a spherical model with a range of 100 m and a sill of 0.3846. The nugget effect is a discontinuity at the origin of the variogram and relates to measurement errors and to spatial sources of variations at distances smaller than shortest sampling interval (Journel and Huijbregts, 1978). The variogram model for the total count (ROI_{TC}) included three basic structures (Fig. 1): (1) a nugget effect of 0.5387, (2) a spherical model with a range of 18 m and a sill of 0.1363; (3) a spherical model with a range of 100 m and a sill of 0.5516.. The variogram model for the soil silt content included two basic structures (Fig. 1): a nugget effect of 0.1964 and a spherical model with a range of 83 m and a sill of 0.9252. The nested variogram model used for the soil water content included three basic structures (Fig. 1): (1) a nugget effect of 0.2619; (2) an exponential model with a practical range of 21 m and a sill of 0.5440; (3) a spherical model with a range of 100 m and a sill of 0.2226. The goodness of fitting was verified by cross validation and the results were quite satisfactory because the statistics used, i.e. mean of the estimation error and variance of the standardised error, were quite close to 0 and 1, respectively.

Finally, using the multi-Gaussian kriging, the spectral regions of interest (ROI) corresponding to potassium-40 (ROI_K) and the total count (ROI_{TC}), the soil water content and the soil silt content, were interpolated and mapped (Fig. 2).

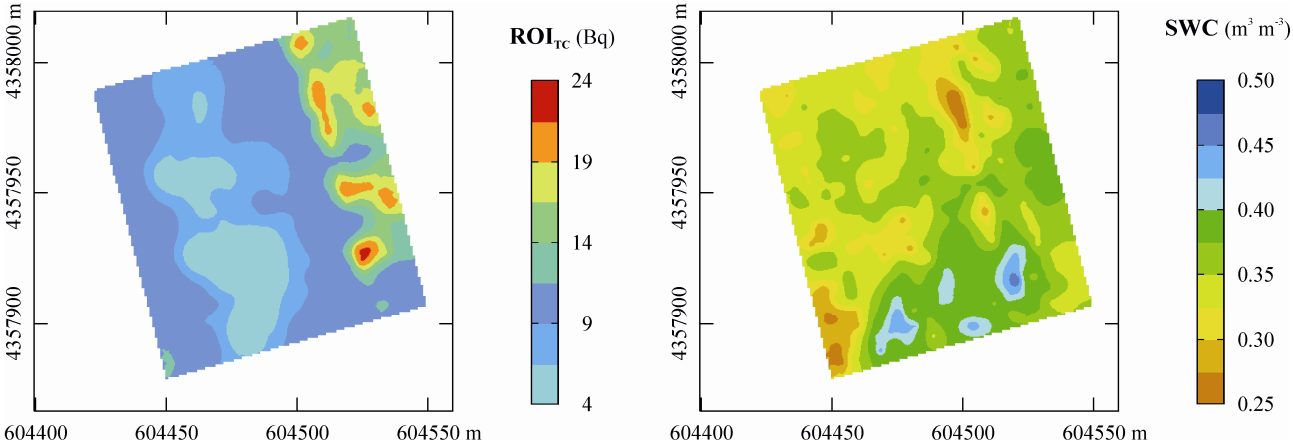


Figure 2: Maps obtained using multi-Gaussian kriging for the spectral regions of interest (ROI) corresponding to the total count (ROI_{TC}) and for the soil water content.

Contrarily to what was expected, there was no clear relation between ROITC and soil water content (Fig. 2). Taylor et al. (2002) reported that the attenuation of gamma-rays through the soil varied with bulk density and water content and the signal attenuation increased by approximately 1% for each 1% increase in volumetric water content (Cook et al., 1996). Probably, the lack of relation was due to the rather homogeneous soil water content (mean content = 35%, standard deviation of 4.7%). A new soil survey with low soil water content will confirm the relation between the activity of radioisotopes and water content.

References

Bihari Á., Dezső Z. 2008. Examination of the effect of particle size on the radionuclide content of soils. Journal of Environmental Radioactivity 99, 1083-1089.

Cook S.E., Corner R.J., Groves P.R., Grealish G.J. 1996. Use of airborne gamma radiometric data for soil mapping. Australian Journal of Soil Research, 34, 183–194.

Dickson B.L., Scott, K.M. 1997. Interpretation of aerial gamma-ray surveys: adding the geochemical factors. AGSO. Journal of Australian Geology and Geophysics, 17, 187–200.

Dowdall, M., O'Dea, J., 2002. Ra-226/U-238 disequilibrium in an upland organic soil exhibiting elevated natural radioactivity. J. Environ. Radioact. 59, 91-104.

Gregory A.F., Horwood, J.L. 1961. A Laboratory Study of Gamma-Ray Spectra at the Surface of Rocks. Mines Branch Research Report R.85. Department of Mines and Technical Surveys, Ottawa.

IAEA, 2003. Guidelines for radioelement mapping using gamma ray spectrometry data. IAEA-TECDOC-1363, International Atomic Energy Agency, Vienna, pp. 173.

Kabata-Pendias, A., Pendias, H., 2001. Trace Elements in Soils and Plants, pp. 413, third ed.. CRC, Boca Raton, Fl.

Navas A., Gaspar L., López-Vicente M., Machín J., 2011. Spatial distribution of natural and artificial radionuclides at the catchment scale (South Central Pyrenees). Radiation Measurements 46, 261-269.

Topp G.C; Davis J.L. 1985. Measurement of soil water content using time-domain reflectometry (TDR): a field evaluation. Soil Science Society of America Journal, 49, 19–24

Verly G. 1983. The multigaussian approach and its application to the estimation of local reserves. Mathematical Geology, 15, 259–286.

Wackernagel H. 2003. Multivariate geostatistics: An introduction with applications. Berlin: Springer-Verlag, p. 387.

Webster, R., Oliver, M. A., 2007. Geostatistics for Environmental Scientists. 2nd Ed. Wiley, Chichester.

Screening level risk assessment for phenols in surface water of three rivers in Tianjin, China [1]

Wenjue ZHONG, Donghong WANG, Zijian WANG *
Research Center for Eco-Environmental Sciences，Chinese Academy of Sciences，
Beijing 100085, e-mail of the corresponding author: wangzj@rcees.ac.cn

Wenjue ZHONG, Lingyan ZHU
College of Environmental Science and Engineering, Nankai University, Key Laboratory of Pollution Processes and Environmental Criteria, Ministry of Education, Tianjin Key Laboratory of Remediation & Pollution Control for Urban Ecological Environment, Tianjin 300071, China

Abstract: The purpose of this paper was to identify the phenols in surface water of three rivers in Tianjin and assess the ecological risk. Using technology of retention time lock (RTL) and deconvolution reporting software (DRS), a contaminants list including all the phenols which were identified in the samples was obtained and all identified phenols were quantified. The concentration levels of total phenols in three rivers accorded with the patterns that Dagu river>Beitang river>Yongdingxinhe river, and June samples>October samples. Risk quotients (RQ) were used to assess the environmental risk of identified phenols. As a result, 5, 6 and 2 phenols were determined as potential ecological risk stressors in surface water of Beitang river, Dagu river and Yongdingxinhe river, respectively.

Keywords: phenols, ecological risk, screening level, DRS

1. Introduction

Phenols exist widely in environment. They can pose many adverse effects to aquatic organisms because of their toxicity, persistence and bioaccumulative potential. In the past decades, many papers have been devoted to the occurrences of phenols in natural waters (House *et al.* 1997; Staples *et al.* 2000; Belfroid *et al.* 2002). However, little information is available for their concentration levels in Chinese rivers. Therefore, it is necessary to screen broad-spectrum phenols in the environment and assess their full-scale ecological risk in order to improve risk control.

Tianjin is the third largest industrial center in China. With intense industrial and commercial activities in the coastal area, rivers in the Tianjin are severely polluted with high loads of persistent organic pollutants and these bring risks to the water environment. Dagu river, Beitang river and Yongdingxinhe river are three main sewage-received rivers. Industrial, agricultural and domestic wastewaters from Tianjin area are directly or indirectly discharged into the three rivers (Song *et al.* 2006). It is important and urgent, therefore, to evaluate the occurrence and ecological risk of phenols in the three rivers.

[1] National Basic Research Program (973) of China (No.2007CB407301); National Natural Science Foundation of China (No.20977102); Important National Science & Technology Specific Projects (2008ZX07314-003-3);

The objectives of this study are to: 1) determine the concentration levels of phenols in surface water of three rivers, consequently, the ecological risks of the idenfied phenols were characterized; 2) select the potential ecological risk stressors as priorities for further ecological risk assessment based on the risk quotients.

2. Materials and Methods

2.1.Chemicals and materials

All the phenolic standards were purchased from Sigma-Aldrich (USA) and the detailed information of phenolic standards were listed in the Table 1.

Compound	Abbr.	Compound	Abbr.	Compound	Abbr.
2,4-dinitrophenol	2,4-DNP	4-Nitrophenol	4-NP	Dichlorophene	
Phenol	-	2,3,5-Trichlorophenol	2,3,5-TCP	Hexanoestrol	
2-cresol	-	2,4,5-Trichlorophenol	2,4,5-TCP	Bithionol	
3-cresol	-	2,3,6-Trichlorophenol	2,3,6-TCP	Hexachlorophene	
4-cresol	-	4-Chlororesorcinol	-	Pyrocatechol	-
2-chlororphenol	2-CP	4-chloro-2-nitrophenol	4-C-2-NP	2-naphthol	-
2,4- Xylenol	-	2-Chlorohydroquinone	-	2-Biphenylol	-
4-chlorophenol	4-CP	3,4,5-Trichlorophenol	3,4,5-TCP	Resorcinol	-
4-Chloro-3-methylphenol	4-C-3-MP	2-Chloro-4-nitrophenol	2-C-4-NP	2-Nitrophenol	2-NP
2,5-Dichlorophenol	2,5-DCP	2,3,5,6-Tetrachlorophenol	2,3,5,6-TeCP	Hydroquinone	-
2,6-dichlorophenol	2,6-DCP	2,3,4,6-Tetrachlorophenol	2,3,4,6-TeCP	2,6-Xylenol	-
2,3,6-Trimethylphenol	2,3,6-TMP	2,3,4,5-Tetrachlorophenol	2,3,4,5-TeCP	2-Isopropylphenol	-
2,4-dichlorophenol	2,4-DCP	2,4-dichloro-3-ethyl-6-nitrophenol	-	2-sec-Butylphenol	-
2,6-Diisopropylphenol	-	Pentachlorophenol	PCP	4,4'-Biphenyldiol	-
p-chloro-m-xylenol	PCMX	Ortho-benzyl-para-chlorophenol	-	Biphenol A	BPA
3,5,6-trichloro-2-pyridinol	-	2-chloro-4-phenylphenol	-	6-chlorothymol	-
2,4,6-Trichlorophenol	2,4,6-TCP	Tetrachlorohydroquinone	-		

--Abbr.: Abbreviation; "-": no abbreviation available

Table 1: The detailed information of 50 phenols

2.2 Sampling and preparation

39 and 31 surface water samples (2 L) were collected in Tianjin, China with aid of a global positioning system (GPS; Fig. 1) in June and October, 2007, respectively. b1-b7 were sampled in Beitang river, d1-d19 were sampled in Dagu river and y1-y15 were sampled in Yongdingxinhe river. The method for sample conservation and preparation and analytical procedures could see in literatures published before(Zhong *et al.* 2010; Zhong *et al.* 2011).

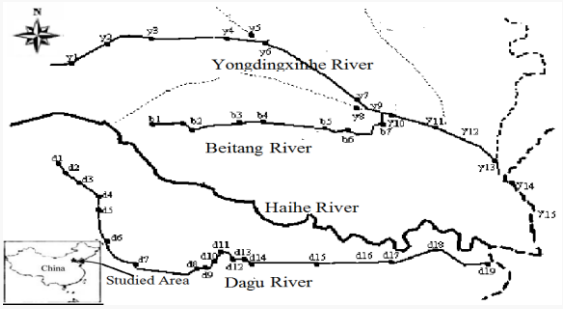


Fig 1: Sampling sites

2.3 Approach for screening level ecological risk assessment

The quotient method was used to characterize risk. Risk quotient (RQ) was defined as the ratio of predict environmental concentration (PEC) and predict no effect concentration (PNEC). Chemicals of potential concerns (COPCs) with RQ exceeding 1.0 were selected as potential stressors.

Using spectral deconvolution technology, a contaminants list including all the phenols which were identified in the samples was obtained. All phenols identified in samples were quantified and their concentration levels were used as PEC.

Chronic Value (Chv) were collected from PBT Profiler (USEPA 2010). As the recommendation of OECD, ten was taken as the assessment factor (AF). So tenth of Chvs were used as PNEC (USEPA 1985; OECD 1995).

3. Results

3.1 Identification and quantification of phenols in surface water of three rivers

The qualitative and quantitative results are listed in Table 2. The concentration levels of total phenols in three rivers accorded with the patterns that Dagu river> Beitang river> Yongdingxinhe river, and June samples> October samples.

chemicals	Beitang river		Dagu river		Yongdingxin river		PNEC
	June	October	June	October	June	October	
Phenol	nd-10.3	nd -10.9	nd -520	nd -15.1	nd -0.1	nd -1.05	19
2- cresol	nd-52.6	nd -15.6	nd -45.3	nd -3.09	nd -1.35	nd -0.2	12
3- cresol	nd-18.7	nd -16.4	nd -386	nd -1.06	nd -0.51	nd -0.11	12
2,4-xylenol	nd-32.8	nd -20.5	nd -90.6	nd -0.33	nd -6.61	nd -3.74	7.8
4-CP	nd-0.44	nd -1.63		nd -0.11		nd -0.05	13
2-CP						nd -0.07	13
2,5-DCP	nd-2.43	nd -4.63	nd -1.23		nd -0.29		8.5
2,6-DCP			nd -0.16				8.5
2,4,6-TCP				nd -0.16	nd -0.32	nd -0.99	5.3
4-NP	nd-1.77			nd -1.75	nd -0.18		18
2,3,6-TMP	nd-1.92	nd -1.35					5
PCMX	nd-3.7	nd -1.36	nd -157				5.2
2-naphthol	0.34-16.4			nd -167	nd -4.58	nd -0.15	8.5
Resorcinol		nd -0.35	nd -0.69		nd -0.17		36
Pyrocatechol			nd -0.04				36
2-Biphenylol	nd-0.51	nd -0.25	nd 0.95	nd -0.09			5.4
2-sec-Butylphenol	nd-11.1	nd -6.59		nd -1.98		nd -0.04	4
2,4-dichloro-3-ethyl-6-nitrophenol	nd-0.99						2.8

--nd: not detected;

Table 2: The results of identifying and quantifying phenols in three rivers and PNEC (μg/L)

3.2 Ecological risk assessment

Chv were collected from PBT Profiler (USEPA 2010) (last accessed February, 2010) and tenth of Chvs were used as PNEC (Table 2).

Using quotient method, phenols whose RQ exceed 1.0 were picked out. These phenols were considered as potential stressors to aquatic environment of three rivers. Furthermore, the risk levels of each potential stressors were sorted by RQs. Five kinds of phenols were selected as priority phenols in Beitang river, and the order of risk level was 2-cresol>2-sec-butylphenol>2,4-xylenol>2-naphthol>3-cresol. Six kinds of phenols were selected as priority phenols in Dagu river, and the order of risk level was 3-

cresol>PCMX>phenol>2-naphthol>2,4-xylenol>2-cresol. For Yongdingxinhe river, all RQs were less than 1.0. Although only COPCs with RQ exceeding 1.0 present a clear risk potential, any COPCs with the quotient greater than 0.3 are subjected to more rigorous risk assessment (WERF 1996), because chemical interactions and cumulative toxicity must also be considered. So 2-naphthol and 2,4-xylenol with RQ > 0.3 were selected as priority phenols for Yongdingxinhe river.

4. Concluding remarks

DRS was used to identify broad-spectrum phenols in three rivers of Tianjin, China. The result indicated that phenols exist widely in three rivers. Using quotient method to assess ecological risk of identified phenols, 5, 6 and 2 phenols were determined as potential ecological risk stressors in surface water of Beitang river, Dagu river and Yongdingxinhe river, respectively.

References

Belfroid A., van Velzen M., van der Horst B. & Vethaak D. (2002). Occurrence of bisphenol A in surface water and uptake in fish: evaluation of field measurements. *Chemosphere*, 49, 97-103.

Dav ìM.L. & Gnudi F. (1999). Phenolic compounds in surface water. *Water Research*, 33, 3213-3219.

House W.A., Leach D., Long J.L.A., Cranwell P., Smith C., Bharwaj L., Meharg A., Ryland G., Orr D.O. & Wright J. (1997). Micro-organic compounds in the Humber rivers. *Science of The Total Environment*, 194-195, 357-371.

OECD (1995). OECD environment monographs 92, Guidance document for aquatic effects assessment. OECD/GD (95) 18.

Song M., Xu Y., Jiang Q., Lam P.K.S., O'Toole D.K., Giesy J.P. & Jiang G. (2006). Measurement of estrogenic activity in sediments from Haihe and Dagu River, China. *Environment International*, 32, 676-681.

Staples C.A., Dorn P.B., Klecka G.M., O'Block S.T., Branson D.R. & Harris L.R. (2000). Bisphenol A concentrations in receiving waters near US manufacturing and processing facilities. *Chemosphere*, 40, 521-525.

USEPA (1985). Guidelines for deriving numerical national water quality criteria for the protection of aquatic organism and their uses. National Technical Information Service Accession Number PB85-227049. US Environmental ProtectionAgency, Washington D C.

USEPA (2010). Environmental Science Center PBT profiler. http://www.pbtprofiler.net/Results.asp (last accessed February, 2010).

WERF (1996). Aquatic ecological risk assessment: a multi-tiered approach. Alexanfria, Virginia

Zhong W.J., Wang D.H., Xu X.W., Luo Q.A., Wang B.Y., Shan X.Q. & Wang Z.J. (2010). Screening level ecological risk assessment for phenols in surface water of the Taihu Lake. *Chemosphere*, 80, 998-1005.

Zhong W.J., Wang D.H., Xu X.W., Wang B.Y., Luo Q.A., Kumaran S.S. & Wang Z.J. (2011). A gas chromatography/mass spectrometry method for the simultaneous analysis of 50 phenols in wastewater using deconvolution technology. *Chinese Science Bulletin*, 56, 275-284.

Spatial Dynamic Factor Models with environmental applications

Pasquale Valentini - Luigi Ippoliti
University G. d'Annunzio (Italy), pvalent@unich.it

Dani Gamerman
Universidade Federal do Rio de Janeiro (Brasil)

Abstract: This article is concerned with a dynamic factor model for spatio-temporal environmental variables. The model is proposed in a state-space formulation which, through the Kalman recursions, allows a unified approach to prediction and estimation. Full probabilistic inference for the model parameters is facilitated by adapting standard Markov chain Monte Carlo (MCMC) algorithms for dynamic linear models to our model formulation.

Keywords: Dynamic factor models, Spatio-temporal models

1 Introduction

In recent years, spatio-temporal models have received widespread popularity and have been largely developed through applications in environmental sciences. In fact, the European Environmental Agency and the US Environmental Protection Agency have both devoted significant efforts to developing air quality models for the assessment of air pollution issues and evaluation of feasible solutions. We note nevertheless that there is no single approach which can be considered uniformly as being the most appropriate for a specific problem.
In this paper, we propose a latent regression model which is useful for spatial and temporal predictions of pollutants of interest. The model is developed in a state-space representation which represents a powerful way to provide full probabilistic inference for the model parameters, interpolation and forecast of the variable of interest. To account for spatial interpolation, the spatial dependence is incorporated in the measurement matrix and we describe its construction by discussing a stochastic specification. The possibility of specifying two measurement equations leads to a significant advantage in terms of spatial interpolation and this makes an important difference with respect to other spatio-temporal models proposed in literature. A further important property of the proposed model is that it leads to capture the temporal variation of the multivariate space-time fields.

2 The General Model

Consider the multivariate spatio-temporal processes $\mathbf{X}(\mathbf{s},t) = [X_1(\mathbf{s},t),\ldots,X_{n_x}(\mathbf{s},t)]'$ and $\mathbf{Y}(\mathbf{s},t) = [Y_1(\mathbf{s},t),\ldots,Y_{n_y}(\mathbf{s},t)]'$, where $\mathbf{s} \in S$, with S a some spatial domain and $t \in \{1,2,\ldots\}$ a discrete index of times. For geostatistical data, S is a given subset of $\Re^d$ and s is assumed to vary continuously throughout S . For lattice data, S is assumed to be a given finite or countable collection of points. Lattices may be either regular, as on a grid, or irregular, such as zip codes, census divisions.
It is explicitly assumed that $\mathbf{X}$ is a predictor of $\mathbf{Y}$. Hence, $\mathbf{Y}$ denotes the specific multivariate process of interest to be predicted in time and/or space. Here, the relationship between the multivariate processes is modelled through the structural spatial dynamic factor (SSDF) model.
Let us define the multivariate spatial processes as $\mathbf{Y}(t) = [\mathbf{Y}(\mathbf{s}_1,t)',\ldots,\mathbf{Y}(\mathbf{s}_N,t)']'$, a $\tilde{n}_y \times 1$ vector ($\tilde{n}_y = n_y N$) at N locations for n_y variables, and $\mathbf{X}(t) = [\mathbf{X}(\mathbf{s}_1,t)',\ldots,\mathbf{X}(\mathbf{s}_N,t)']'$, a $\tilde{n}_x \times 1$ vector ($\tilde{n}_x = n_x N$) at N locations for n_x variables.
The measurement equations of the SSDF model are

$$\begin{aligned} \mathbf{X}(t) &= \mathbf{m}_x(t) + \mathbf{H}_x\mathbf{f}(t) + \mathbf{u}_x(t) \quad (1)\\ \mathbf{Y}(t) &= \mathbf{m}_y(t) + \mathbf{H}_y\mathbf{g}(t) + \mathbf{u}_y(t) \quad (2)\end{aligned}$$

where $\mathbf{m}_y(t)$ and $\mathbf{m}_x(t)$ are $\tilde{n}_y \times 1$ and $\tilde{n}_x \times 1$ mean components modelling the smooth large-scale temporal variability, $\mathbf{H}_y$ ($\tilde{n}_y \times m$) and $\mathbf{H}_x$ ($\tilde{n}_x \times l$) are measurement matrices giving information on the spatial structure of the random fields, and $\mathbf{u}_x(t) \sim N(\mathbf{0},\boldsymbol{\Sigma}_{u_x})$ and $\mathbf{u}_y(t) \sim N(\mathbf{0},\boldsymbol{\Sigma}_{u_y})$. Throughout the paper it is assumed that $m \ll \tilde{n}_y$ and $l \ll \tilde{n}_x$.
The temporal dynamic of the processes is modelled through the following state equations:

$$\begin{aligned} \mathbf{g}(t) &= \sum_{i=1}^{p} \mathbf{B}_i\mathbf{g}(t-i) + \sum_{j=1}^{q} \mathbf{C}_j\mathbf{f}(t-j) + \boldsymbol{\xi}(t) \quad (3)\\ \mathbf{f}(t) &= \sum_{k=1}^{s} \mathbf{R}_k\mathbf{f}(t-k) + \boldsymbol{\eta}(t) \quad (4)\end{aligned}$$

where $\mathbf{C}_i$ ($m \times m$), $\mathbf{D}_j$ ($m \times l$), and $\mathbf{R}_k$ ($l \times l$) are coefficient matrices modelling the temporal evolution of the latent vectors $\mathbf{g}(t) = [g_1(t),\ldots,g_m(t)]'$ and $\mathbf{f}(t) = [f_1(t),\ldots,f_l(t)]'$, respectively. Finally, $\boldsymbol{\xi}(t)$ and $\boldsymbol{\eta}(t)$ are independent Gaussian error terms for which we assume, $\boldsymbol{\xi}(t) \sim N(\mathbf{0},\boldsymbol{\Sigma}_\xi)$ and $\boldsymbol{\eta}(t) \sim N(\mathbf{0},\boldsymbol{\Sigma}_\eta)$.
SSDF analysis may be used to identify possible clusters of locations whose temporal behaviour is primarily described by a potentially small set of common dynamic latent factors.

3 The Structural Spatial Dynamic Factor Model

It is customary for dynamic factor models to refer to the unobserved (state) processes as the common factors and to refer to the coefficients that link the factors with the observed series as the factor loadings. Because of their spatial nature, the factor loadings are equivalently defined as spatial patterns (Lopes et al., 2008; Ippoliti et al, 2010). The latent factors, $\mathbf{f}(t)$ and $\mathbf{g}(t)$, are able to capture the temporal variation of the multivariate space-time fields, and the spatial dependence can be modeled by the columns of the matrices $\mathbf{H}_y$ and $\mathbf{H}_x$ through multivariate Gaussian Random Field for geostatistical data, or through multivariate Markov random field (MRF) for lattice data.

4 The State Space Formulation

Given the SSDF model the temporal dynamic is modelled through state equations (3) and (4). The specification of equation (4) is necessary to predict in time the latent process $\mathbf{f}(t)$ and thus to obtain $k-$step ahead forecasts of $\mathbf{g}(t)$ through equation (3). The joint generation process of $\mathbf{g}(t)$ and $\mathbf{f}(t)$ is a VAR(p) process of the type

$$\mathbf{d}(t) \;=\; \mathbf{\Phi}_1\mathbf{d}(t-1) + \ldots + \mathbf{\Phi}_p\mathbf{d}(t-p) + \boldsymbol{\epsilon}(t) \tag{5}$$

where
$\mathbf{d}(t) = \begin{bmatrix} \mathbf{g}(t) \\ \mathbf{f}(t) \end{bmatrix}$, $\mathbf{\Phi}_i = \begin{bmatrix} \mathbf{C}_i & \mathbf{D}_i \\ \mathbf{0} & \mathbf{R}_i \end{bmatrix}$, $\boldsymbol{\epsilon}(t) = \begin{bmatrix} \boldsymbol{\xi}(t) \\ \boldsymbol{\eta}(t) \end{bmatrix}$ and $p \geq max(s,q)$.
The presence of the measurement and the state variables naturally leads to the state-space representation of the SSDF model.

5 Nonstationary case

In the case in which the two spatio-temporal processes $X(\mathbf{s};t)$ and $Y(\mathbf{s};t)$ are not stationary in time, we assume that factors are generated by cointegrated vector autoregressive processes. In this case the factors are represented by the error correction specification of the vector autoregressive process of equation (5):

$$\Delta\mathbf{d}(t) \;=\; \tilde{\mathbf{A}}\mathbf{d}(t-1) + \sum_{i=1}^{p-1} \tilde{\mathbf{\Phi}}_i \Delta\mathbf{d}(t-i) + \boldsymbol{\epsilon}(t) \tag{6}$$

where $\tilde{\mathbf{A}} = -\mathbf{I} + \sum_{i=1}^{p} \mathbf{\Phi}_i$, $\tilde{\mathbf{\Phi}}_i = -\sum_{j=i+1}^{p} \mathbf{\Phi}_l$, and Δ is the difference operator (i.e. $\Delta\mathbf{d}(t) = \mathbf{d}(t) - \mathbf{d}(t-1)$).
Let $\mathbf{\Phi}(z)$ denote the characteristic polynomial associated with the process (6). We assume that latent exogenous variables are cointegrated with cointegrating rank r_f

(Cho, 2010) and also $rank(\tilde{\mathbf{A}}) = r$, $r = m + l - c > r_f$ with $m + l > c$ and c are the unit roots of $Det(\mathbf{\Phi}(z))$.
Because of the exogeneity of X, the matrices $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{\Phi}}_i$ are upper block triangular matrices: $\tilde{\mathbf{A}} = \begin{bmatrix} \tilde{\mathbf{A}}_1 & \tilde{\mathbf{A}}_{12} \\ \mathbf{0} & \tilde{\mathbf{A}}_2 \end{bmatrix}$ and $\tilde{\mathbf{\Phi}}_i = \begin{bmatrix} \tilde{\mathbf{\Phi}}_{1i} & \tilde{\mathbf{\Phi}}_{12i} \\ \mathbf{0} & \tilde{\mathbf{\Phi}}_{2i} \end{bmatrix}$.
Then, equation (6) can be rewritten in the following two equations:

$$\Delta \mathbf{g}(t) = \mathbf{A}\mathbf{B}'\mathbf{d}(t-1) + \mathbf{A}_{2f}\mathbf{B}_f'\mathbf{f}(t-1) + \sum_{i=1}^{p-1} \mathbf{K}_i \Delta \mathbf{d}(t-i) + \boldsymbol{\xi}(t) \quad (7)$$

$$\Delta \mathbf{f}(t) = \mathbf{A}_f\mathbf{B}_f'\mathbf{f}(t-1) + \sum_{i=1}^{p-1} \tilde{\mathbf{\Phi}}_{2i} \Delta \mathbf{f}(t-j) + \boldsymbol{\eta}(t) \quad (8)$$

where $\mathbf{A}$ is $m \times r_d$, $\mathbf{B}$ is $(m + l) \times r_d$, $\mathbf{A}_f$ and $\mathbf{B}_f$ are $l \times r_f$, $\mathbf{A}_{2f}$ is $m \times r_f$, $\mathbf{K}_i = [\tilde{\mathbf{\Phi}}_{1i} \;\; \tilde{\mathbf{\Phi}}_{12i}]$ and $r_d \leq m + l$.

6 Inference

Full probabilistic inference for the model parameters is carried out by elicitating the independent prior distributions. Posterior inference for the proposed class of spatial dynamic factor models is facilitated by MCMC algorithms. The common factors are jointly sampled by means of the well known forward filtering backward sampling (FFBS) algorithm (Carter and Kohn 1994) which exploits the state space representation of the general model. All other full conditional distributions are "standard" multivariate normal distributions or gamma distributions. An exception is for the spatial parameters and the covariance matrices which are sampled using a Metropolis-Hastings step.

References

Carter C. K., Kohn R. (1994) On Gibbs sampling for state space models. *Biometrika*, 81, 541-553.

Cho S. (2010), Inference of cointegrated model with exogenous variables, *SIRFE Working Paper* 10-A04, Seoul National University.

Ippoliti L., Valentini P., Gamerman D. (2010) Space-time modelling of coupled spatio-temporal environmental variables, *Technical Report* N. 229, DME/IM-UFRJ.

Lopes H. F., Salazar E., Gamerman D. (2008) Spatial dynamic factor analysis. *Bayesian Analysis*, 3, 759-792.

Spatial Point Processes Applied to the Study of Forest Fires in Portugal[1]

Paula Sequeira Pereira
ESTSetubal-IPS/CEAUL, paula.pereira@estsetubal.ips.pt

Kamil Feridun Turkman
DEIO-FCUL/CEAUL, kfturkman@fc.ul.pt

Abstract: The aim of this work is to analyse the behaviour of forest fires in Portugal using statistical techniques applied to spatial point processes. We present a short overview on the most commonly used summary statistics for spatial point processes under homogeneity and inhomogeneity assumptions. The data set consist of records of 6295 forest fires larger than 100 hectares, observed in Portugal during the years 1975 through 2005.

Keywords: spatial point process, K-function, marks

1 Introduction

Forest fires are a major environmental problem in Portugal. In the past few years thousands of hectares of forest have been destroyed. The aim of this work is to analyse the behaviour of forest fires in Portugal using statistical techniques applied to spatial point processes. With this analysis we intend to investigate whether the forest fires occur randomly, in clusters or in some regular pattern and we examine if the marked spatial point pattern depends on forest fires size.

This analyse is conducted in the context of a preliminary analysis of forest fires in Portugal, which is part of the general objective of modelling the location and the forest fires sizes by an adequate marked spatio-temporal point processes.

2 Materials and Methods

The point pattern under investigation consists of satellite imagery records of 6295 forest fires larger than 100 hectares, observed in Portugal during the years 1975 through 2005, acquired annually after the end of the summer fire season.

A conventional starting point for the analysis of a spatial point process is to investigate the hypothesis of complete spatial randomness (CSR). A process is CSR when we have a homogeneous Poisson point process, i.e. the intensity is constant,

[1]The research was supported by FCT/OE projects and SFRH/PROTEC/67394/2010 PhD grant.

and the events are independents of each other and have the same propensity to be found at any locations. If the CSR hypothesis is rejected, then must be a tendency towards clustering (events occur in closely spaced groups) or regularity (events more spaced than under CSR).

A popular tool to describe departures from CSR is the Ripley's K-function, $K(r) = \lambda^{-1}E[\text{number of events within distance } r \text{ of an arbitrary event}]$, where λ is the intensity of the point process (the number of events per unit area). Under CSR, the Ripley's K-function is simply $K(r) = \pi r^2$. Comparing the shape of ours Ripley's K-function relative to the shape of the Ripley's K-function in the case of CSR provides valuable information on the point process distribution. A Ripley's K-function that deviates from CSR can indicate that events interact or have some effect on each other, but it can also indicate that exists a trend in the pattern (the intensity of the process must not be constant across the region). By using an inhomogeneous K-function to analyse the data, is removed the assumption of an underlying homogeneous point process. The inhomogeneous K-function has the same interpretation as the homogeneous Ripley's K-function, except that the intensity of events is no longer constant but depends on the locations of the events.

In general, is common to use the L-function, which is defined as $L(r) = \sqrt{K(r)/\pi}$. Under CSR, $L(r) = r$. So we can use the line through the origin as a reference and it is simple to detect clustering or regularity by graphing $L(r) - r$ against the distance r (as in Figure 2).

The fires are characterized not only by their position but also by its size (area burned), which can be interpreted as a mark. So in addition to detection of clustering among points, the relationship between marks and between points and marks are investigated with mark correlation function defined by Stoyan (Illian et al (2008)) and **E** and **V** functions defined by Schlather et al (2004). The aim of the mark correlation function is to find out whether the marks are correlated and the aim of the **E** and **V** functions is to find out whether marks and locations are correlated (whether marks depend on local point density). If this last hypothesis is not rejected then the process is simplified greatly because the point pattern and the marks could be investigated separately.

We used the statistical software `R` and functions in the spatial point processes library `spatstat` to compute the various results.

3 Results

We start the analysis of the forest fires in a purely spatial context, so we drawn a map of the locations of the forest fires and the corresponding map of the intensity for the whole period, 1975 to 2005. Figure 1 shows that larger forest fires occur in the centre and south of Portugal but the majority of the fires are in the north of Portugal. Almost in whole country the fires are less than 0.05 fires per km^2, but in the north the highest values of the intensity are achieved, 0.2 fires per km^2.

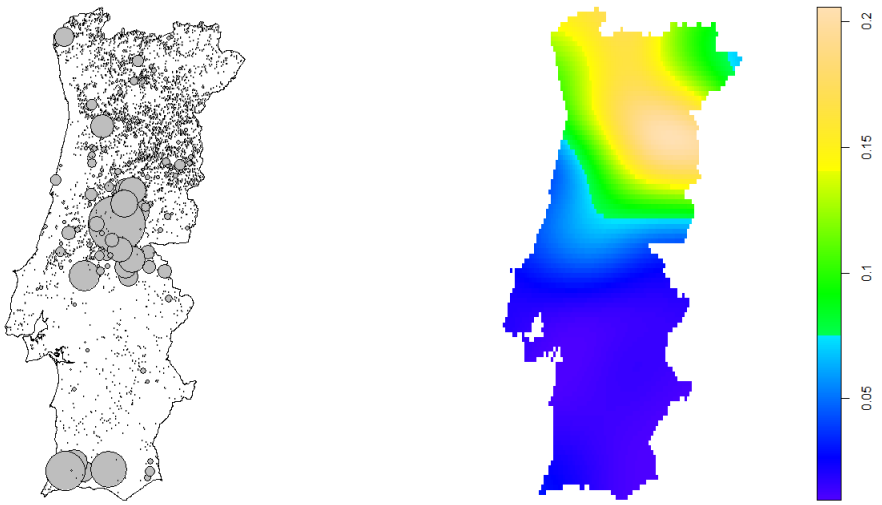


Figure 1: *Left*: Point plot of the locations of the forest fires (the circles are proportional to the area burned); *Right*: Kernel estimation of the intensity function

We compute the homogeneous and inhomogeneous L-function for all the period, 1975 to 2005. As we can see in Figure 2, there is a clear departure from CSR towards clustering. But the homogeneous L-function is overestimating the amount of clustering present in the point pattern. When we compute the inhomogeneous L-function$-r$ the distance between the estimated L-function$-r$ and the upper envelope reduces a lot. However, the inhomogeneous L-function shows evidence of clustering with a radius approximately of 50 km.

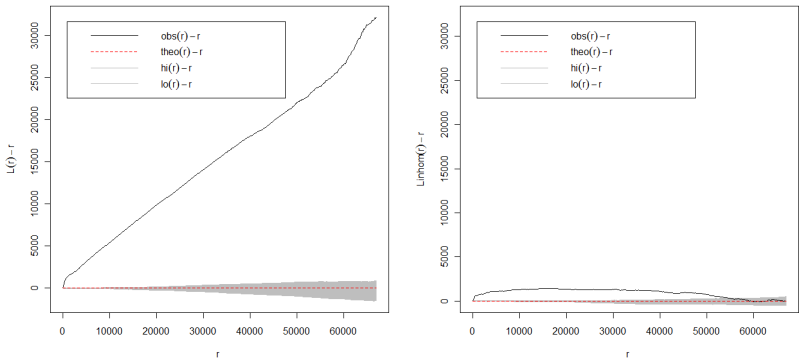


Figure 2: *Left*: Homogeneous L-function$-r$; *Right*: Inhomogeneous L-function$-r$. (In all graphics pointwise envelope under CSR and r in meters)

The mark correlation function for the area burned of the forest fires is shown in Figure 3. The shape of the empirical mark correlation function reveals that the marks do not appear to be correlated. The **E** and **V** functions indicate that the

model does not appear to belong to the random field model, i.e., does not appear to be independent of the unmarked point process.

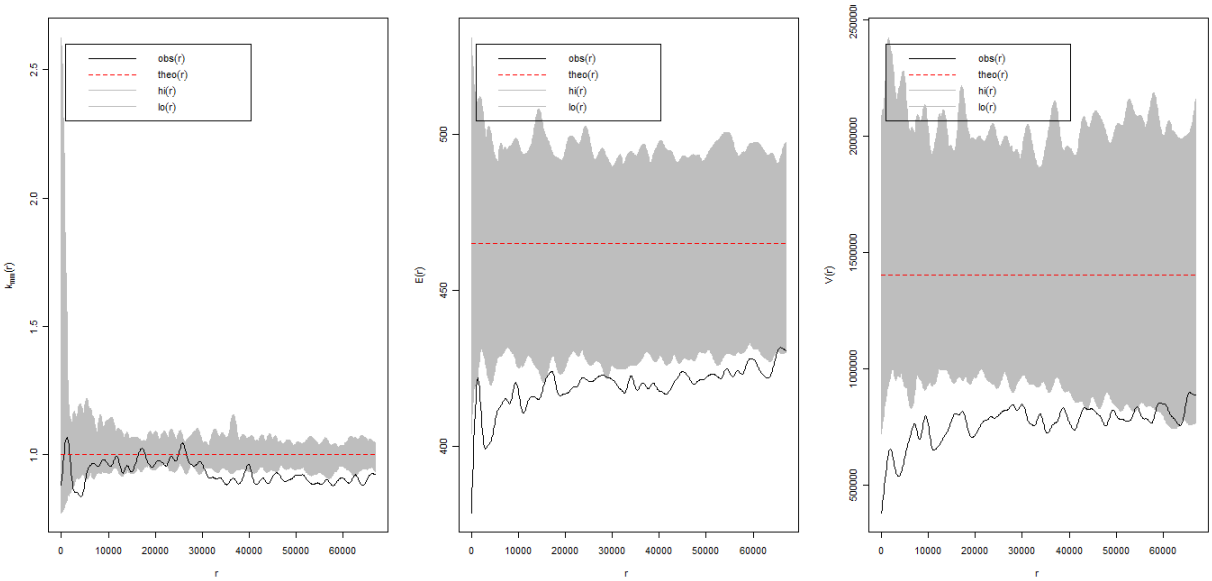


Figure 3: *Left*: Mark correlation function; *Centre*: **E** function; *Right*: **V** function. (In all graphics pointwise envelope under CSR and r in meters)

4 Concluding remarks

From the modelling point of view, the information about the behaviour of homogeneous and inhomogeneous L-function is important because it suggests that trend should be included in a model for forest fires and show that forest fires events occur in clusters, indicating spatial dependence or interaction.

The marks, area burned, do not appear to be correlated. The hypothesis of the marks to be a random field independent of the unmarked point process appears to be rejected. The point pattern and the marks should not be investigated separately.

References

Illian J., Penttinen A., Stoyan H., Stoyan D. (2008) *Statistical analysis and modelling of spatial point patterns*, Wiley. Statistics in Practice.

Schlather M., Ribeiro P., Diggle P. (2004) Detecting dependence between marks and locations of marked point processes. *J. R. Statist. Soc., Ser. B*, 66, 79-93.

Spatio-temporal Analysis of Forest Fires in Portugal [1]

Maria Inês Dias
Dep. Mathematics and CIMA, Universidade de Évora, Portugal, misd@uevora.pt

Giovani Loiola da Silva
Dep. Mathematics - IST, Technical University of Lisbon and CEAUL, Portugal

Abstract: In the last decade, forest fires have become one of the worst natural disasters in Portugal, causing great forest devastation, leading to both economic and environmental losses and putting at risk populations and the livelihoods of the forest itself. In this paper we present Bayesian hierarchical models to analyze spatio-temporal fire data on the proportion of burned area in Portugal, by municipalities and over three decades. Mixture of distributions was employed to model jointly the proportion of area burned and the excess of no burned area for early years. For getting estimates of the model parameters, we used Monte Carlo Markov chain methods.

Keywords: Forest fires, spatio-temporal data, Bayesian hierarchical models.

1 Introduction

According to the National Forestry Authority (*Direcção Geral dos Recursos Florestais*), Portugal has the largest number of fires among five Mediterranean countries (Portugal, Spain, France, Italy and Greece), being important to look for spatio-temporal patterns of fires e.g. modeling the proportion of burned area. As the proportion of burned area (Y) is a continuous variable and restricted to the interval (0, 1), we can model it by assuming naturally a beta distribution (Ferrari and Cribari-Neto, 2004) or Gaussian distribution and a Skew-Normal (Azzalini and Dalla Valle, 1996) distributions after a *logit* transformation, *i.e.* $\log(Y/(1-Y))$. In addition, we can use Bayesian hierarchical models to take into account spatially correlated random effects (Silva *et al.*, 2008) and excess zeros in the proportion of burnt area by municipalities and years (Amaral-Turkman *et al.*, 2010). Our aim is to present a spatio-temporal analysis of forest fires in mainland Portugal, by 278 municipalities between 1980 and 2006, from a Bayesian point-of-view and using Monte Carlo Markov chain (MCMC) methods to obtain estimates of the parameters of interest.

[1] This work is partially supported by FCT.

2 Materials and Methods

Let Y_{it} the proportion of burned area in municipality i and year t, $i = 1, \ldots, n$, $t = 1, \ldots, T$. Assume Y_{it} or $\log(Y_{it}/(1-Y_{it}))$ has a probability distribution with mean μ_{it} and variance σ^2. Silva *et al.* (2008) suggest that μ_{it} can be expressed by $\mu_{it} = \alpha + S_0(t) + S_i(t) + \phi_i$, where $S_0(t)$ can represent a nonlinear temporal effect, $S_i(t)$ is the temporal effect by region i and ϕ_i a random effect of the spatial variation associated with region i. If $\phi_i = b_i + h_i$, component h_i represents the unstructured spatial random effect with Gaussian priori distribution ($h_i \sim N(0, \sigma_h^2 \equiv \frac{1}{\tau_h})$, and b_i the spatially correlated random effect with priori distribution, $p(b_i|\tau_b = \frac{1}{\sigma_b^2})$, chosen in terms of a conditional autoregressive model (CAR) (Besag *et al.*, 1981), *i.e.*, $b_i \mid \mathbf{b}_{-i}, \sigma_b^2 \sim N(\bar{b}_i, \sigma_b^2/m_i)$, where $\bar{b}_i$ is the mean of the random effects related to the "neighbors" of the region i, m_i the number of adjacent regions to region i and σ_b^2 the variance component.

Upon the occurrence of zeros, the distribution of the proportion of area burned (Y_{it} is considered a mixture of distributions with probability function $f(y_{it})$, denoting $f_1(y_{it}) = f(y_{it}|y_{it} \neq 0)$, $i = 1, \ldots, n$, $t = 1, \ldots, T$. Define V_{it} as a Bernoulli random variable such that, $V_{it} = 0$, with probability p_{it_0}, and 1, with probability $p_{it_1} \equiv 1 - p_{it_0}$, where p_{it_0} represents the probability of non-burned area in the region i in the year t. V_{it} indicates the existence of the burnt area in the region i in the year t. Thus, $f(y_{it}) = f_1(y_{it})^{V_{it}} (1 - p_{it_0})^{V_{it}} p_{it_0}^{1-V_{it}}$. The probability of no burned area in the region i at time t is modeled as, $\log(\frac{p_{it_0}}{1-p_{it_0}}) = \beta_0 + \beta_1 t + \psi_i$, where ψ_i is a CAR model. We use assigned highly dispersed but proper priors. In fact, one typically assumes independent normal prior for the regression coefficients. For the variance component hyperparameters, one usually assigns an inverse gamma prior, *e.g.*, $\sigma^2 \sim IG(r_1, s_1)$, $\sigma_b^2 \sim (r_2, s_2)$, $\sigma_h^2 \sim IG(r_3, s_3)$ and $\sigma_\psi^2 \sim IG(r_4, s_4)$ with kernel density given for $x^{-(r+1)} exp(-s/x)$, $x > 0$. Consequently, we can construct the related joint posteriori distribution and use MCMC methods because the corresponding marginal posteriors are not easy to get explicitly. Notice that these methods are implemented *e.g.* in WinBUGS (Spiegelhalter *et al.*, 2007).

3 Results and Concluding remarks

Based on the models in Section 2, we analyze the proportion Y_{it} of burnt area due to forest fires in 278 municipalities (mainland Portugal) and over 27 years (1980-2006). Data were collected by Portuguese National Forestry Authority. Three scenarios were considered for the data modeling:

A) Gaussian probability model: $logit(Y) \sim N(\mu, \sigma^2)$;

B) Skew-normal model: $logit(Y) \sim SN(\mu, \sigma^2, \lambda)$, where λ is a shape parameter;

C) Beta model: $Y \sim Beta(a, b)$, with $E[Y] = \mu$, $Var(Y) = \frac{\mu(1-\mu)}{\gamma+1}$ and $\gamma = a+b$.

By using MCMC methods via WinBUGS, we used 15,000 iterations for all fitted models, taking every 10th iteration of the simulated sequence, after 5000 iterations of burn-in. In Table 1, one can be observed some fitted models and, based on the Deviance Information Criterion (DIC), the selected model is model M_4. Note that $S_0(t) = \eta_t$, in model M_4, represents a second order random walk. For selected model

Model		p_D	DIC ($\times 10^6$)
$M_1(A)$	$\mu_{it} = \beta_0 + \beta_1 t + \phi_i t + b_i + h_i$ $logit(p_{it}) = \delta_0 + \delta_1 t + a_i$	521	150.150
$M_2(B)$	$\mu_{it} = \beta_0 + \beta_i t + \phi_i t + b_i + h_i$ $logit(p_{it}) = \delta_0 + \delta_1 t + a_i$	509	150.150
$M_3(C)$	$logit(\mu_{it}) = \beta_0 + \beta_1 t + \phi_i t + b_i + h_i$ $logit(p_{it}) = \delta_0 + \delta_1 t + a_i$	581	149.996
$M_4(C)$	$logit(\mu_{it}) = \beta_0 + \eta_t + b_i$ $logit(p_{it}) = \delta_0 + \delta_1 t + a_i$	411	149.995

Table 1: Model selection based on DIC

(M_4), the posteriori mean, standard deviation (s.d.) and 95% higgest posterior density (HPD) credible Intervals (CI) of some parameters of interest are in Table 2. Based on model M_4, the spatio-temporal risks of burned area, defined here by $exp(\eta_t + b_i)$ for municipality i, were used to produce maps in 1985, 1994 and 2001 (Figure 1), as well as maps for spatial risks $\exp(b_i)$ and $\exp(a_i)$ (Figure 2).

parameter	mean	s.d.	95% CI
δ_1	-0.169	0.007	(-0.183, -0.156)
γ	24.82	0.449	(24.02, 25.69)
σ_b^2	0.334	0.051	(0.237, 0.437)
σ_η^2	3.357	0.508	(2.424, 4.379)
σ_a^2	0.194	0.060	(0.098, 0.313)
p_{it_0}	0.143	0.003	(0.137, 0.150)

Table 2: Estimates of the model parameters (M_4)

The spatio-temporal analysis of the burned area proportion in 278 municipalities of mainland Portugal between 1980 and 2006 reveals an increasing trend in the proportion of burned area, whereas the number of municipalities without burned area trend to decrease. The space-time models studied here have smoothed estimates used in the production of maps that are useful in the interpretation of spatio-temporal data. This analysis of the Portuguese forest fires may isolate trends in small areas of administrative knowledge for promoting an appropriate policy interventions to reduce that national catastrophe.

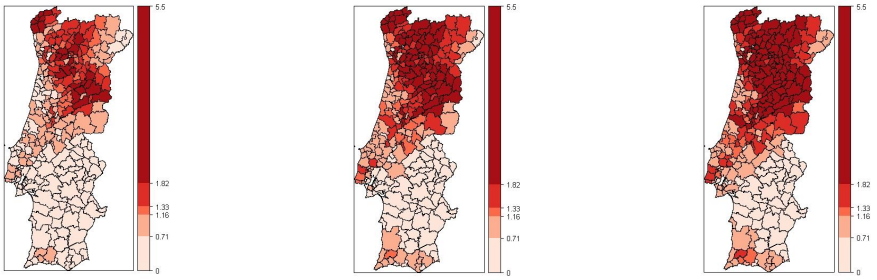


Figure 1: Spatio-temporal risks in 1985 (left), 1994 (middle) and 2001 (right)

Figure 2: Spatial risk maps - $\exp(b_i)$ (left) and $\exp(a_i)$ (right)

References

Amaral-Turkman, M.A., Turkman, K.F., Le Page. Y., Pereira, J.M.C. (2011) Hierarchical space-time models for fire ignition and percentage of land burned by wildfires. *Environ. Ecol. Stat.* DOI 10.1007/s10651-010-0153-9.

Azzalini, A., Dalla Valle A. (1996). The Multivariate Skew-normal Distribution. *Biometrika* 83, 715-726.

Besag, J., York, J.C., Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* 43, 159.

Ferrari, S.L.P., Cribari-Neto, F.(2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics* 31, 799–815.

Silva, G.L., Dean, C.B, Niyonsenga, T. and Vanasse, A. (2008). Hierarchical Bayesian spatiotemporal analysis of revascularization odds using smoothing splines. *Statistics in Medicine* 27, 2381-2401.

Spiegelhalter, D., Thomas, A., Best, N.G., Lunn, D. (2007). *Bayesian inference using Gibbs sampling for Windows (WinBUGS), Version 1.4.3.* Cambridge: MRC Biostatistics Unit (http:\\www.mrc-bsu.cam.ac.uk\bugs\).

SPATIAL2
Spatial Data Methods for Enviromental and Ecological Processes - 2nd Edition

Thursday, September 1, 2011 - Morning - Faculty of Law, University of Foggia

08:15-09:30 Registration (Aula Magna)

09:30-10:00 Welcome and introductory remarks (Aula Magna)

10:00-10:30 Coffee break (Sala Consiglio)

10:30-11:30 Plenary session

P1 - Aula Magna - Sampling and Accurate Predictions for Environmental Management

Chair A. Pollice, University of Bari

- **Variograms to Guide Spatial Sampling for Kriging, M.A. Oliver, R. Kerry**
- **Generalised Kriging with Environmental Applications, L. Ippoliti**

11:30-12:30 Specialized sessions

S1 - Room #2

Air Quality - Chair L. Malherbe, INERIS

- **Application of a modeling system aimed at studying the impact on air quality of a waste storage fire, Giua R., Morabito A., Tanzarella A.**
- **Estimation of the areas of air quality limit value exceedances on national and local scales. A geostatistical approach, Malherbe L., Beauchamp M., Létinois L., Ung A., de Fouquet C.**
- **Modeling pollutant threshold exceedance probabilities in the presence of exogenous variables, Ignaccolo R., Sylvan D., Cameletti M.**
- **Using the SPDE approach for air quality mapping in Piemonte region, Cameletti M., Lindgren F., Simpson D., Rue H.**

S2 - Room #4

Animal and Plant Ecology - Chair P. Monestiez, INRA

- **A generalization of the Incidence Function Model for metapopulations with fluctuating behaviour: an application to Lymantria dispar (L.) in Sardinia, Bodini A., Gilioli G., Cocco A., Lentini A., Luciano P.**
- **Geostatistical modelling of regional bird species richness: exploring environmental proxies for conservation purpose, Bacaro G., Chiarucci A., Santi E., Rocchini D., Pezzo F., Puglisi L.**
- **Spatial Bayesian Modelling of Presence-only Data, Divino F., Golini N., Jona Lasinio G., Penttinen A.**
- **The deep-water rose shrimp in the Ionian Sea: a spatio-temporal analysis of zero-inflated abundance data, D'Onghia G., Maiorano P., Carlucci R., Tursi A, Pollice A., Ribecco N., Calculli C., Arcuti S.**

S3 - Room #5

Methods and Environmental Modeling - Chair C. Miller, University of Glasgow

- **Applying a new procedure for fitting a multivariate space-time linear coregionalization model, De Iaco S., Palma M., Posa D.**
- **Decision making for root disease control: a problem in reducing the nugget variance, Correll R.**
- **EM estimation of the Dynamic Coregionalization Model with varying coefficients, Finazzi F., Fassò A.**
- **Likelihood Inference in Multivariate Model-Based Geostatistics, Ferrari C., Minozzo M.**

S4 - Room #6

Proximal and Remote Sensing in Precision Agriculture - Chair A. Castrignanò, CRA-SCA of Bari

- **A system for on-line measurement of key soil properties, Mouazen A.M., Kuang B., Quraishi M.Z.**
- **Multimodal remote sensing for enhancing detection of spatial variability in agricultural fields, Alchanatis V., Cohen A., Cohen Y., Levi O., Naor A.**
- **Modified Hot-Spot analysis for spatio-temporal data: a case study of the leaf-roll virus expansion in vineyards, Cohen Y., Sharon R., Sokolsky T., Zahavi T.**
- **The use of the geoadditive model with interaction in a Precision Agriculture context: a comparison of different spatial correlation structures, Cafarelli B., Crocetta C., Castrignanò A.**

12:30-14:00 Transfer from Foggia to Baia delle Zagare

Thursday, September 1, 2011 - Afternoon - Baia delle Zagare

14:00-16:30 Lunch and accommodation

16:30-17:30 Specialized sessions

S5 - Room #1

Landscape Ecology and Natural Resource Management

Chair F. Bruno, University of Bologna

- **Comparison of spatial statistics for identifying underlying process in forest ecology, Brown C., Illian J., Burslem D., Law R.**
- **Connectivity in a real fragmented landscape: distance vs movement model based approaches, Mairota P., Leronni V., Cafarelli B., Baveco J.M.**
- **Methodological study on pesticides in Alsatian groundwater, Musci F., Giasi C.I., de Fouquet C.**
- **The GIS approach to detect the influence of the fresh water inflows on the marine-coastal waters: the case of the Apulia Region (Italy) through standard monitoring data, Porfido A., Barbone E., La Ghezza V., Costantino G., Perrino V., Ungaro N., Blonda M.**

S6 - Sala Conferenze

Space-Time Surveillance for Public Health

Chair A. Biggeri, University of Firenze

- **Modeling malaria incidence in Sucre state, Venezuela, using a Bayesian approach, Villalta D., Guenni L., Rubio Y.**
- **Prediction of cancer mortality risks in spatio-temporal disease mapping, Goicoa T., Ugarte M.D., Militino A.F., Etxeberria J.**
- **Predictive assessment of a non-linear random effects model for space-time surveillance data, Paul M., Held L.**
- **Selective Inference in Disease Mapping, Catelan D., Biggeri A.**

17:30-18:30 Plenary session

P2 - Sala Conferenze

Ecology and Water Analysis

Chair G. Jona Lasinio, Sapienza University of Roma

- **Assessing Temporal and Spatial Change in Nutrients for Large Hydrological Areas, C. Miller, A. Magdalina, A.W. Bowman, E.M. Scott, D. Lee, R. Willows, C. Burgess, L. Pope, D. Johnson**
- **Definition of type-specific reference conditions in Mediterranean lagoons, A. Basset, E. Barbone, I. Rosati**

18:30-19:30 Posters and drinks

20:00 Dinner

SPATIAL2
Spatial Data Methods for Enviromental and Ecological Processes - 2nd Edition

Friday, September 2, 2011 - Morning – Baia delle Zagare

09:00-10:00 Specialized sessions

S7 - Room #1
Environmental Data Analysis
Chair T. Gneiting, University of Heidelberg

- **A software for optimal information based downsizing/upsizing of existing monitoring networks, Barca E., Passarella G., Vurro M., Morea A.**
- **Comparing SaTScan and Seg-DBSCAN methods in spatial phenomena, Montrone S., Perchinunno P., L'Abbate S., Ligorio C.**
- **Fire, earthquake, landslide, volcano, flood: first approach to a natural hazard map of Italy, Camporese R., Iandelli N.**
- **Spatio-Temporal Analysis of Wildfire Patterns in Galicia (NW Spain), Fuentes-Santos I., Gonzalez-Manteiga W., Marey-Pérez. M.F.**

S8 - Sala Conferenze
Sampling Designs for Natural Studies
Chair D. Cocchi, University of Bologna

- **On the design-based properties of spatial interpolation, Bruno F., Cocchi D., Vagheggini A.**
- **Relations between spatial design criteria, Mueller W.G., Waldl H.**
- **Simulation-based optimal design for estimating weed density in agricultural fields, Bel L., Parent E., Makowski D.**
- **The dramatic effect of preferential sampling of spatial data on variance estimates, Clifford D., Kuhnert P., Dobbie M., Baldock J., McKenzie N., Harch B., Wheeler I., McBratney A.**

10:00-11:00 Plenary session

P3 - Sala Conferenze

Ensemble Forecasts

Chair L. Bel, AgroParisTech

- **Ensemble forecasting: status and perspectives, F. Nerozzi, T. Diomede, C. Marsigli, A. Montani, T. Paccagnella**
- **Statistical postprocessing for ensembles of numerical weather prediction models, T. Gneiting**

11:00-11:30 Coffee break

11:30-12:30 Specialized sessions

S9 - Sala Conferenze
Climatology and Meteorology
J. Mateu, Universitat Jaume I

- **A few links between the notion of Entropy and Extreme Value Theory in the context of analyzing climate extremes, Naveau P., Rietsch T., Guillou A., Merleau J.**
- **Geoadditive modeling for extreme rainfall data, Bocci C., Petrucci A., Caporali E.**
- **Spatio-temporal rainfall trends in southwest Western Australia, Liang K., Chandler R., Marra G.**
- **Stochastic Downscaling of Precipitation with Conditional Mixture Models, Carreau J, Vrac M.**

S10 – Room #1
Space-Time Surveillance of Natural Assets
Chair C. Crocetta, University of Foggia

- **Geostatistical modeling of ice content within the "Glacier Bonnard" (Switzerland), Jeannee N., Faucheux C., Bardou E., Ornstein P.**
- **Is space-time interaction real or apparent in seismic activity?, Rotondi R., Varini E.**
- **Spatio-temporal modelling for avalanche risk assessment in the North of Italy, Nicolis O, Assuncao R.**
- **A seismic swarm as a dynamic ergodic stochastic process: a case study of the L'Aquila's earthquake in 2009, Coli M.**

12:30-15:00 Lunch

Friday, September 2, 2011 – Afternoon – Baia delle Zagare

15:00-16:00 Specialized sessions

S11 - Sala Conferenze

Disease Mapping and Environmental Exposure - Chair Ignaccolo R., University of Torino

- **A Bayesian Spatio-Temporal framework to improve exposure measurements combining observed and numerical model output, Pirani M., Gulliver J., Blangiardo M.**
- **A spatio-temporal model for cancer incidence data with zero-inflation, Musio M., Sauleau E.A.**
- **Generalized Estimating Equations for Zero-Inflated Spatial Count Data, Monod A.**
- **Poisson M-Quantile Geographically Weighted Regression on Disease mapping, Chambers R., Dreassi E., Salvati N.**

S12 - Room #1

GIS and Soil Sciences - Chair B. Cafarelli, University of Foggia

- **Imputation strategy in spatial data, Martino L., Palmieri A.**
- **Multivariate geostatistical model to map soil properties at a region scale from airborne hyperspectral imagery and scattered soil field surveys: dealing with large dimensions, Monestiez P., Walker E., Gomez C., Lagacherie P.**
- **Optimal location and size for a biomass plant: application of a GIS methodology to the "Capitanata" district, Monteleone M., Cammerino A.R.B., lo Storto M.C.**
- **Population Density in a City, Abbate C., Salvucci G.**

16:00-17:00 Plenary session

P4 - Sala Conferenze

Climatology and Meteorology - Chair P. Naveau, LSCE - CNRS

- **Global temperature analysis with non-stationary random field models, F. Lindgren, H. Rue, P. Guttorp**
- **Methods for climate change detection and attribution, A. Ribes**

17:00-18:00 Spatial Café with poster discussion - Organizer C. Crocetta

Table #1 - Agriculture, biodiversity, groundwater pollution and hydrogeology - Facilitators: A. Castrignanò and Y. Cohen

Table #2 - Air quality and disease mapping - Facilitators: P. Dawid and A. Pollice

Table #3 - Climatology and meteorology and sampling design - Facilitators: W. Mueller and A. Petrucci

Table #4 - Ecology, conservation and natural resources management - Facilitators: C. de Fouquet and G. Jona Lasinio

Table #5 - Environmental risk assessment - Facilitators: D. Cocchi and L. Guenni

18:00-19:00 Plenary Session

P5 - Sala Conferenze - Spatial Functional Data - Chair A. Fassò, University of Bergamo

- **Spatially correlated functional data, J. Mateu**
- **Clustering of environmental functional data, A. Pastore, S. Tonellato, R. Pastres**

19:00-19:30 Concluding remarks

20:00 Social dinner

Saturday, September 3, 2011 – Morning – Baia delle Zagare

9:15 Transfer from Baia delle Zagare to Foggia

9:00-13:00 INLA Tutorial: "Fast Bayesian inference for Geostatistics and other latent Gaussian models", F. Lindgren

13:30 ONLY FOR THOSE ATTENDING THE INLA TUTORIAL Transfer from Baia delle Zagare to Foggia

Spatial Café - Organizer C. Crocetta

Table #1 - Agriculture, biodiversity, groundwater pollution and hydrogeology - Facilitators: A. Castrignanò and Y. Cohen

- A data driven model for spatio-temporal estimation of shallow water table depth in soils, Ungaro F., Calzolari C.
- A Methodology for Evaluating the Temporal Stability of Spatial Patterns of Vineyard Variation, Gambella F., Dau R., Paschino F., Castrignanò A., De Benedetto D.
- Assessment and modelling of spatial variability of the soil factors potentially affecting groundwater nitrate contamination in two agricultural areas of Molise Region (Southern Italy), Colombo C., Palumbo G., Sollitto D., Castrignanò A.
- Assessment of Spatial and Temporal Within-Field Soil Variability by Using Geostatistical Techniques, Castrignanò A., Cucci G., Diacono M., De Benedetto D., Lacolla G., Troccoli A.
- CYCAS-MED project: analysis at regional and local scale of climate change impacts on cereals yield in Morocco, Bodini A., Entrade E., Cesaraccio C., Duce P., Zara P., Dubrovsky M.
- Geostatistical analysis and mapping of hydrocarbon pollutants in soils, de Fouquet Chantal
- Geostatistical analysis of groundwater nitrates distribution in the Plaine d'Alsace, Spacagna R.L., De Fouquet C., Russo G.
- Influence of different olive grove management on spider diversity, Loverre P., Addante R, Calculli C.
- Landcover classification of agricultural sites using multi-temporal COSMO-Skymed data, Satalino G., Balenzano A., Belmonte A., Mattia F., Impedovo D.
- Multidimensional analysis of data from Bari Harbour: a GIS based tool for the characterization and management of bottom sediments, Dellino P., Mele D., Mega M., Pagnotta E., De Giosa F., Taccardi G., Ungaro N., Costantino G.
- Multivariate statistical analyses for the source apportionment of groundwater pollutants in Apulian agricultural sites, Ielpo P., Cassano D., Lopez A., Abbruzzese De Napoli P., Pappagallo G, Uricchio V.F.
- Structural changes in seismic activity before large earthquakes, Gallucci M., Petrucci A.
- Using environmental metrics to describe the spatial and temporal evolution of landscape structure and soil hydrology and fertility, Pascual Aguilar J. A., Sanz Garcia J., de Bustamante Gutierrez I., Kallache M.

Spatial Café - Organizer C. Crocetta

Table #2 - Air quality and disease mapping - Facilitators: P. Dawid and A. Pollice

- A comparison between hierarchical spatio-temporal models in presence of spatial homogeneous groups: the case of Ozone in the Emilia-Romagna Region, Bruno F., Paci L.
- A multilevel multimember model for smoothing a disease map of lung cancer rates, Bartolomeo N., Trerotoli P., Serio G.
- A spatio-temporal model for air quality mapping using uncertain covariates, Cameletti M., Ghigo S., Ignaccolo R.
- African dust contribution on the PM10 daily exceedances occurred in Apulia region, Angiuli L., Giua R., Loguercio Polosa S., Morabito A.
- Health impact assessment of pollution from incinerator in Modugno (Bari), Galise I, Serinelli M., Bisceglia L., Assennato G.
- Local scoring rules for spatial processes, Dawid P., Musio M.
- Measuring Urban Quality of Life Using Multivariate Geostatistical Models, Michelangeli A., Ferrari C., Minozzo M.
- Multivariate and Spatial Extremes for the Analysis of Air Quality Data, Padoan S., Fassò A.
- Pulmonary Tuberculosis and HIV/AIDS in Portugal: joint spatio-temporal clustering under an epidemiological perspective, Nunes C., Briz T., Gomes D., Filipe P.A.
- Spatial diffusion and temporal evolution of PCDD/Fs, PCBs and PAHs congener concentrations in the ambient air of Taranto: an analysis based on the duality diagram approach, Pollice A., Esposito V.
- Spatial disaggregation of pollutant concentration data, Horabik J., Nahorski Z.
- Spatial representativeness of an air quality monitoring station. Application to NO2 in urban area, Beauchamp M., Malherbe L., Létinois L., de Fouquet C.
- Statistical investigations on PAH concentrations at industrial sampling site, Amodio M., Andriani E., Dambruso P.R., de Gennaro G., Demarinis Loiotile A., Di Gilio A., Trizio L., Assennato G., Colucci C., Esposito V., Giua R., Menegotto M, Spartera M.
- Tapering spatio temporal models, Fassò A., Finazzi F., Bevilacqua M.

Spatial Café - Organizer C. Crocetta

Table #3 - Climatology and meteorology and sampling design - Facilitators: W. Mueller and A. Petrucci

- **Alternative approaches for probabilistic precipitation forecasting, Bruno F., Cocchi D., Rigazio A.**
- **Comparison of Calibration Techniques for Limited-Area Ensemble Precipitation Forecast Using Reforecasts, Diomede T., Marsigli C., Montani A., Paccagnella T.**
- **Functional boxplots for summarizing and detecting changes in environmental data coming from sensors, Romano E., Balzanella A., Rivoli L.**
- **Information, advice, friendship, notes and trust network: evidence on learning from classmate, Zavarrone Emma, Vitali Agnese**
- **Optimal spatial design for air quality measurement surveys: what criteria?, Romary T., de Fouquet C., Malherbe L.**
- **Point-process statistical analysis for the ECMWF Ensemble Prediction System, Nerozzi F.**

Spatial Café - Organizer C. Crocetta

Table #4 - Ecology, conservation and natural resources management - Facilitators: C. de Fouquet and G. Jona Lasinio

- Combining geostatistics and process-based water quality model to improve estimation along a stream network. Example on a stretch of the Seine River, de Fouquet C., Polus-Lefèbvre E., Flipo N., Poulin M.
- Landscape impacts of photovoltaic plants on the ground: a case-study through the application of rendering techniques, Robles N., Primerano R., Perrino V., Blonda M.
- Marine spatial planning in Apulia (Italy): Reconciling seagrass conservation with the multiple use of coastal areas, Fraschetti S., Lembo G., Tursi A., D'Ambrosio P., Terlizzi A., De Leo F., Paes S., Guarnieri G., Bevilacqua S., Boero F.
- Regional estimation method of rivers low flow from river basin characteristics, Rossi G., Caporali E.
- Spatial Analysis of some soil physicochemical properties in mountainous massif of Sico, Portugal, Torres O. M., Neves M. M., Gomes D. P.
- Spatial and auto corrrelation of ecological change: disturbance and perturbation analysis in Circeo National Park (south Latium, Italy), Galante G., Cotroneo R., Mandrone S., Strafella I.
- Spatial diversity in a "zoom-lens": Analysing ecological communities through weighted spatial scales, Studeny A. C., Brown C., Illian J.B.
- Spatio-temporal changes of biodiversity indices in the bathyal demersal assemblages of the Ionian Sea, Maiorano P., Giove A., Minerva M., Sion L., D'Onghia G , Pollice A., Ribecco N., Muschitiello C.
- Spatio-temporal variability in stream flow status: Candelaro river case study, De Girolamo A.M., Calabrese A., Pappagallo G., Santese G., Lo Porto A., Gallart F, Prat N., Froebrich J.
- Statistical assessment of the plant protection level within protected areas (PA) based on remote sensing products, Menconi M.E., Pacicco C.L.
- Statistical calibration of the Carlit index in the Pontine Island of Zannone, Jona Lasinio G., Tullio M.A., Abdelahad N., Scepi E., Sirago S., Pollice A.
- Statistical issues in the assessment of urban sprawl indices, Cocchi D., Altieri L., Scott M., Ventrucci M., Pezzi G.
- Using spatial statistics tools on remote-sensing data to identify fire regime linked with savanna vegetation degradation, Jacquin A., Chéret V., Goulard M., Sheeren D.

Spatial Café - Organizer C. Crocetta

Table #5 - Environmental risk assessment - Facilitators: D. Cocchi and L. Guenni

- A methodology for assessing the spatial distribution of static wildfire risk over wide areas: the case studies of Liguria and Sardinia (Italy), Bodini A., Entrade E., Cossu Q. A., Canu S., Fiorucci P., Gaetani F., Paroli U.
- A new procedure for fitting a multivariate space-time linear coregionalization model, De Iaco S., Palma M., Posa D.
- Bayesian hierarchical models: An analysis of Portugal road accident data, Ribeiro C., Turkman A. A., Cardoso J.L.
- Electrical Resistivity Measurements for Spatial Soil Moisture Variability Estimation, Calamita G., Luongo R., Perrone A., Lapenna V., Piscitelli S., Straface S.
- Geostatistics and GIS: tools for environmental risk assessment, Maggio S., Cappello C., Pellegrino
- How to estimate anisotropic attenuation exploiting prior isotropic knowledge, Rotondi R., Zonno G.
- Natural radioactivity distribution and soil properties: a case study in southern Italy, Guagliardi I., Ricca N., Cipriani M.G., Civitelli D., Froio R., Gabriele A.L., Buttafuoco G., De Rosa R.
- Screening level risk assessment for phenols in surface water of three rivers in Tianjin, China, Zhong W., Wang D., Wang Z., Zhu L.
- Spatial Dynamic Factor Models with environmental applications, Valentini P., Ippoliti L. , Gamerman D.
- Spatial Point Processes Applied to the Study of Forest Fires in Portugal, Pereira P.S., Turkman K.F.
- Spatio-Temporal Analysis of Forest Fires in Portugal, Dias M. I., da Silva G.L.

SPATIAL2

Spatial Data Methods for Environmental and Ecological Processes - 2nd Edition

1/2 September 2011

Foggia - Baia delle Zagare - ITALY

The 2011 European Regional Conference of The International Environmetrics Society. Satellite of the 58th World Statistics Congress of the International Statistical Institute (ISI).

CONSIGLIO PER LA RICERCA E LA SPERIMENTAZIONE IN AGRICOLTURA

CRA-SCA
UNITA' DI RICERCA PER I SISTEMI CULTURALI DEGLI AMBIENTI CALDO-ARIDI

ARPA PUGLIA

Scientific Committee
Liliane Bel, Agroparistech - France
Barbara Cafarelli, University of Foggia - Italy
Annamaria Castrignanò, CRA of Bari - Italy
Daniela Cocchi, University of Bologna - Italy, President
Corrado Crocetta, University of Foggia - Italy
Alessandro Fassò, University of Bergamo - Italy
Giovanna Jona Lasinio, Sapienza University of Roma - Italy
Alessio Pollice, University of Bari - Italy
Marian Scott, University of Glasgow – United Kingdom

Organizing Committee
Barbara Angelillis, University of Foggia - Italy
Francesca Bruno, University of Bologna - Italy
Barbara Cafarelli, University of Foggia - Italy, President
Giacoma Girone, CRA of Bari - Italy
Rosalba Ignaccolo, University of Torino - Italy
Giovanna Jona Lasinio, Sapienza University of Roma - Italy
Alessio Pollice, University of Bari - Italy
Alessia Spada, University of Foggia - Italy

The International ENVIRONMETRICS Society - TIES

SPATIAL2

Spatial Data Methods for Environmental and Ecological Processes - 2nd Edition

PROCEEDINGS

EDITOR: Barbara Cafarelli

The International ENVIRONMETRICS Society - TIES

ISBN 978-88-96025-12-3

CDP SERVICE EDIZIONI