Jointly modelling air quality and meteorological variables using the D-STEM software

Crescenza Calculli[1], Annarita Turnone[2], Francesco Finazzi[3,*], Alessio Pollice[4] and Alessandro Fassò[5]

[1] *Italian Institute for Nuclear Physics, INFN - Bari, via E. Orabona n. 4, 70125 Bari, Italy;*
calculli.enza@gmail.com
[2] *Apulia Region Environmental Protection Agency (ARPA Puglia), corso Trieste n. 27, 70126 Bari, Italy;*
a.turnone@arpa.puglia.it
[3] *Dept. of Management, Economics and Quantitative Methods, University of Bergamo, via dei Caniana n.2, 24127 Bergamo, Italy;*
francesco.finazzi@unibg.it
[4] *Dept. of Economics and Mathematical Methods, University of Bari, largo Abbazia S. Scolastica n.53, 70124 Bari, Italy;*
alessio.pollice@uniba.it
[5] *Dept. of Engineering, University of Bergamo, viale Marconi n.5, 24044 Dalmine (BG), Italy;*
alessandro.fasso@unibg.it
[*] *Corresponding author*

Abstract. *This work discusses a general framework for mapping pollutant concentrations over a region of interest considering both concentration and meteorological data. The framework is based on a multivariate space-time model able to handle missing data and non-colocated monitoring networks. The model is implemented and estimated by the D-STEM (Distributed Space Time Expectation Maximization) software. As a case study, the model is used to produce daily maps of particulate matters and nitrogen dioxide concentrations for the Italian region Apulia. The official data from the air quality monitoring network of Apulia and the meteorological data from the Meteorological Service of Apulia are used, for a total of* 8 *variables and thus a* 8*-variate model. In order to deal with the computational complexity, the Bari INFN high performance grid computing infrastructure is used.*

Keywords. *EM algorithm; Heterogeneous networks; Linear coregionalization model*

1 Introduction

Mapping the airborne pollutant concentration over a region of interest is a classic problem when dealing with air quality assessment (see for instance [2]). In this work, we discuss a general framework for daily mapping multiple pollutants starting from data collected by multiple monitoring networks. The data include both the pollutant concentration measurements and the measurements of some meteorological variables that may drive the pollutant concentration. The problem is made more difficult by the fact that, in general, the air quality network and the meteorological network may differ in the number of stations and their spatial locations. Moreover, each variable can be observed over a subset of network stations and missing data are possible. The problem is solved considering a general multivariate space-time model able to accommodate for the above data characteristics and to estimate the pollutant concentration at unmonitored spatial locations. The model is implemented and estimated by the D-STEM software (see [1]) which is based on the Expectation Maximization algorithm and is suitable for handling large datasets. As new data are made available every day, model parameters are estimated in an adaptive manner, that is, the model is re-estimated every day considering only the data from the previous *n* days. This allows to reduce the computation burden implied by a large *n* and to avoid the influence of "old" data on present estimations. Within this framework, maps of particulate matters (PM_{10}) and nitrogen dioxide (NO_2) concentrations, uncertainty included, are daily provided for the Apulia region, Italy. The rest of the paper discusses the statistical model, the data and some preliminary results.

2 Statistical model

Let $\mathbf{y}(\mathbf{s},t)$ be the q-variate observation vector at spatial location $\mathbf{s}$ and time t. The general model has the following form:

$$\mathbf{y}(\mathbf{s},t) = \boldsymbol{\mu}(\mathbf{s},t) + \boldsymbol{\omega}(\mathbf{s},t) + \boldsymbol{\varepsilon}(\mathbf{s},t) \tag{1}$$

where $\boldsymbol{\mu}(\mathbf{s},t) = \mathbf{X}_{\boldsymbol{\beta}}(\mathbf{s},t)\boldsymbol{\beta}$ is the fixed effect, $\boldsymbol{\omega}(\mathbf{s},t)$ is the random effect given by

$$\boldsymbol{\omega}(\mathbf{s},t) = \sum_{j=1}^{c} \boldsymbol{\alpha}_j \odot \mathbf{x}_j(\mathbf{s},t) \odot \mathbf{w}_j(\mathbf{s},t) + \mathbf{X}_{\mathbf{z}}(\mathbf{s},t)\mathbf{z}(t) \tag{2}$$

while $\boldsymbol{\varepsilon}(\mathbf{s},t)$ is a random error uncorrelated over space and time. The symbol $\odot$ represents the Hadamard product and

$$\begin{aligned} \mathbf{X}_{\boldsymbol{\beta}}(\mathbf{s},t) &= blockdiag(\mathbf{x}_{\boldsymbol{\beta},1}(\mathbf{s},t),...,\mathbf{x}_{\boldsymbol{\beta},q}(\mathbf{s},t)), \\ \mathbf{X}_{\mathbf{z}}(\mathbf{s},t) &= blockdiag(\mathbf{x}_{\mathbf{z},1}(\mathbf{s},t),...,\mathbf{x}_{\mathbf{z},p}(\mathbf{s},t)), \end{aligned}$$

where *blockdiag* is the block diagonal building operator. The vectors of loading coefficients $\mathbf{x}_{\boldsymbol{\beta},i}(\mathbf{s},t)$ have dimensions $1 \times b_i$ for $i = 1,...,q$ while the vectors $\mathbf{x}_{\mathbf{z},k}(\mathbf{s},t)$ have dimensions $1 \times a_k$ for $k = 1,...,p$.

In Equation 2, each multivariate spatial latent variable $\mathbf{w}_j(\mathbf{s},t)$, for each fixed t, is modelled as a linear coregionalization model with the following spatial variance-covariance matrix functions

$$\boldsymbol{\Gamma}_j(\|\mathbf{s}-\mathbf{s}'\|) = \mathbf{V}_j\rho(\|\mathbf{s}-\mathbf{s}'\|;\theta_j,\nu), \tag{3}$$

where $\mathbf{V}_j$ is a valid $q \times q$ correlation matrix and $j = 1,...,c$. On the other hand, the p-dimensional latent component $\mathbf{z}(t)$ has the following Markovian dynamics:

$$\mathbf{z}(t) = \mathbf{G}\mathbf{z}(t-1) + \boldsymbol{\eta}(t)$$

with transition matrix $\mathbf{G}$ assumed to have eigenvalues smaller than 1 in absolute value and innovations $\boldsymbol{\eta}(t) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$. Finally, the elements of $\boldsymbol{\varepsilon}(\mathbf{s},t)$ are independent and normally distributed with variances σ_i^2, $i = 1,...,q$.

The model parameter set is $\boldsymbol{\psi} = \{\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \boldsymbol{\theta}, \mathbf{v}, \mathbf{G}, \mathbf{v}_{\boldsymbol{\eta}}\}$, where $\boldsymbol{\sigma}^2 = (\sigma_1^2,...,\sigma_q^2)^\top$, $\boldsymbol{\alpha}$ is the $cq \times 1$ dimensional vector obtained by stacking $\boldsymbol{\alpha}_1,...,\boldsymbol{\alpha}_c$ and $\mathbf{v}$ is the $cq(q-1)/2 \times 1$ dimensional vector obtained by stacking the unique and non diagonal elements of $\mathbf{V}_1,...,\mathbf{V}_c$.

3 Data sets

Pollutant concentration and meteorological data are provided by the Apulia Region Environmental Protection Agency and by the Associazione Regionale dei Consorzi di Difesa della Puglia, respectively. In this work, we considered a summer season (1^{st} July 2012 - 30^{th} September 2012) characterized by high temperature and an intrusion of long range transported air mass from the African continent, with a large number of sites exceeding the daily PM_{10} limit value of 50 $\mu g/m^3$. PM_{10} and NO_2 are monitored at 73 and 68 stations, respectively. This implies that some stations only measure one pollutant. The missing data rate for the above temporal period is 13%. On the other hand, meteorological variables are measured at 58 stations with no missing data and they are: atmospheric pressure, relative humidity, temperature, wind speed, east-west wind component and north-south wind component. Some covariates are also considered and in particular population density, land elevation, latitude and longitude. Note that, as the air quality and meteorological networks are not colocated (see Figure 1), the meteorological variables are not available at the spatial locations where pollutants are measured. To avoid interpolation, air quality and meteorological variables are all considered as response variable (thus, $q = 8$).

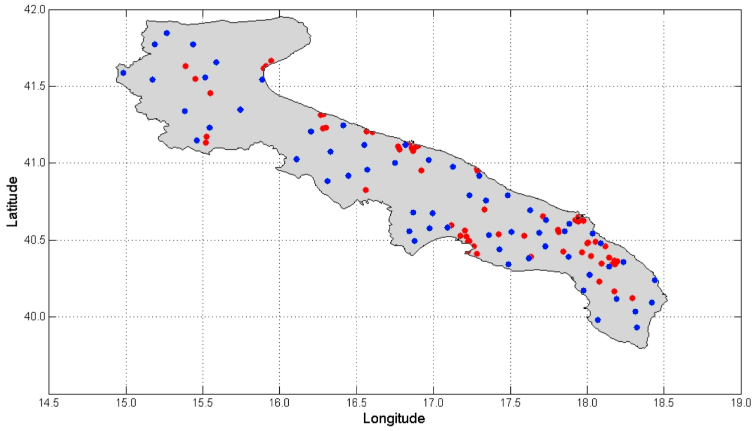

Figure 1: Apulia region air quality network (red dots) and meteorological network (blue dots)

4 Model estimation and mapping

For each day t of the temporal period defined in the previous section, the model in Equation 1 is estimated considering the last 20 days (from $t-19$ to t) and it is eventually used to map the pollutant concentration at day t. This way, model parameters are updated every day and the "old" data are forgotten. This approach is particularly useful if the temporal period is long or if the model is used to produce maps every day as soon as new data are available. As the model is 8-variate, the number of model parameters to be estimated is very high. Nonetheless, the D-STEM software is based on the EM algorithm which is stable and reaches convergence even when q is high. Model estimation and mapping were performed on the Bari INFN high performance grid computing infrastructure. The maps of Figure 2 show the daily average pollutant concentrations (and respective standard deviations) for a particular day within the temporal period.

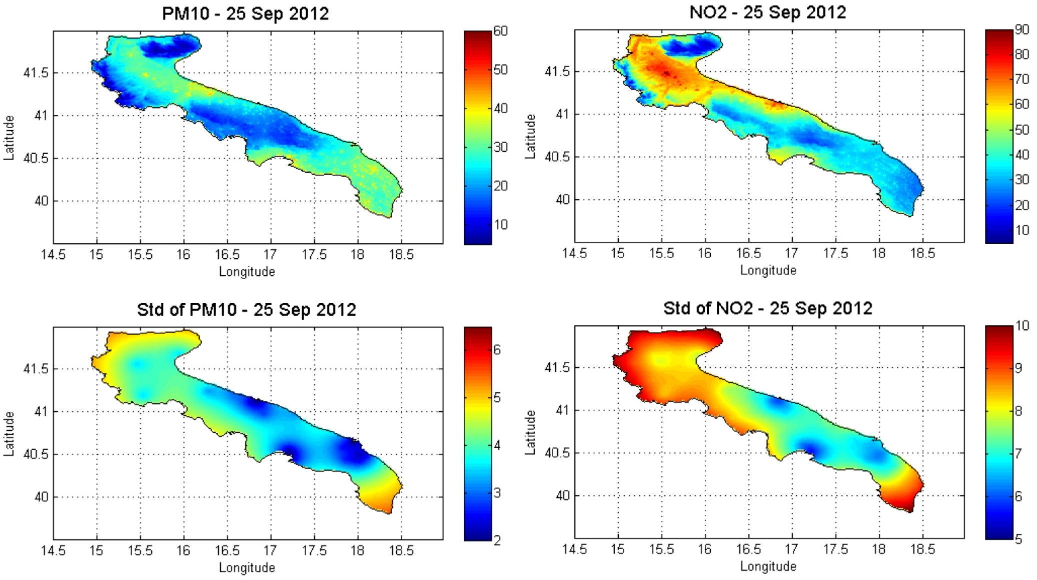

Figure 2: Daily average pollutant concentration and standard deviation maps for PM_{10} and NO_2 on September 25^{th}, 2012

Acknowledgments. This work was partially funded by the FIRB2012 project "Statistical modelling of environmental phenomena: pollution, meteorology, health and their interactions" (RBFR12URQJ). The authors thanks the Associazione Regionale dei Consorzi di Difesa della Puglia, ASSOCODIPUGLIA, for providing the meteorological data and UNIBA-INFN Bari for the access to the high performance grid computing infrastructure ReCas (PON Ricerca e Competitività 2007-2013, Avviso 254/Ric.)

References

[1] Finazzi, F. and Fassò A. (2014). D-STEM: A software for the analysis and mapping of environmental space-time variables. *Journal of Statistical Software*. Accepted.

[2] Finazzi, F., Scott, E. M. and Fassò, A. (2013). A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data. *Journal Of The Royal Statistical Society: Series C*, **62**(2), 287–308.