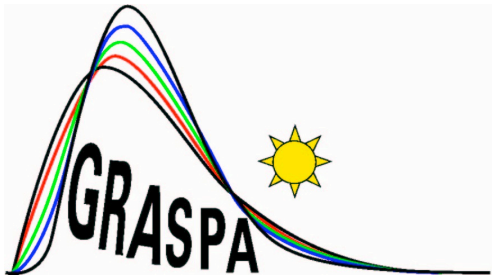
Web Working Papers
by
The Italian Group of Environmental Statistics



Gruppo di Ricerca per le Applicazioni della Statistica
ai Problemi Ambientali

www.graspa.org

A multivariate approach to the analysis of air quality in a high environmental risk area

Alessio Pollice and Giovanna Jona Lasinio

GRASPA Working paper n.32, January 2009

A multivariate approach to the analysis of air quality in a high environmental risk area

Alessio Pollice
Dipartimento di Scienze Statistiche "Carlo Cecchi"
Università degli Studi di Bari

Giovanna Jona Lasinio
Dipartimento di Statistica, Probabilità e Statistiche applicate
Università di Roma "La Sapienza"

Abstract: This study analyzes air quality data in the Taranto municipal area. This is a high environmental risk region being characterized by the massive presence of industrial sites with elevated environmental impact activities. We focus on three pollutants formed by combustion processes and related to meteorological conditions, namely PM10, SO2 and NO2. Preliminary analysis involved addressing several data problems. First of all an imputation technique was considered to cope with the large number of missing data. Missing data imputation was addressed by a leave-one-out procedure based on the recursive Bayesian estimation and prediction of spatial linear mixed effects models enriched by a time-recursive prior structure. Secondly a unique daily weather database at the city level was obtained combining data from 3 stations, characterized by gaps and unreliable measurements. Spatio-temporal modeling of the multivariate normalized daily pollution data was then performed within

a Bayesian hierarchical framework, including time varying weather covariates and a semi-parametric spatial covariance structure. Daily estimates of the pollutants' concentration surfaces allow to identify areas of higher concentration (hot spots), possibly related to specific anthropic activities.

1 Introduction

An analysis of air quality data is provided for the municipal area of Taranto, characterized by high environmental risks due to the massive presence of industrial sites with environmental impacting activities along the NW boundary of the city conurbation. Such activities include iron production (one of the largest plants in Europe), oil-refinery, cement production, fuel storage, power production, waste materials management, mining industry and many others. Some more environmental impacting activities are more deeply integrated within the urban area and have to do with the presence of a large commercial harbour and quite a few military plants (a NATO base, an old arsenal and fuel and munitions storages). These activities have effects on the environment and on public health, as a number of epidemiological researches concerning this area reconfirm (Biggeri et al., 2004). In the context of an agreement between Dipartimento di Scienze Statistiche - Università degli Studi di Bari and ARPA Puglia, air quality data for the municipal area of the city of Taranto were provided. Pollutants continuously monitored by the stations include sulphur dioxide (SO2), nitrogen oxide (NOx) and nitrogen dioxide (NO2), carbon monoxide (CO), benzene, PM10 and ozone.

CO is a toxic gas emitted as a result of combustion processes which, in urban areas, are almost entirely from road traffic emissions, as is NO2. SO2, a corrosive acid gas, is primarily caused by power stations burning fossil fuels which contain sulphur. The present study is focused on PM10, SO2 and NO2 concentrations, temporally and spatially correlated, due to the processes by which the three pollutants are formed. Combustion processes, and diesel combustion in

particular, are a major souce of the pollutants considered here. Slight associations between the levels of pollutants are also observed because of their relationship with meteorological conditions, such as wind direction and speed and temperature. Table 1 shows the sample correlations between the normalized pollutants; notice that the associations between the pollutants are relatively constant across sites (not shown).

At present validated data for years 2005-2007 are available for only 6 monitoring stations managed by the regional government, all equipped with analogous instruments either reporting hourly, two-hourly or daily measurements. Hourly observations of several meteorological variables (including temperature, relative humidity, pressure, rain, solar radiation, wind speed and direction) are also available for 3 weather monitoring stations. Our main objective is to integrate pollution and meteorological data in order to summarize the behaviour of pollution diffusion processes over the area of the municipality for the study period (1st January 2005 - 31 December 2007).

Preliminary data analysis involved addressing quite a few data problems: first we obtained a homogeneous time scale for all monitoring stations transforming the data into daily averages. Normalizing transformations were then applied to the data in order to reach approximate marginal Gaussianity: the square roots of the logs of SO2 and the logs of PM10 and NO2 daily averages were considered. In Tab. 2 a summary of the missing data situation is reported. Missing data are due to both different operational periods of the stations (*staircase missingness*) and occasional malfunction of the sensors (*sparse missing data*); as a consequence an adequate choice between different missing data imputation strategies was required. Finally available weather data are characterized by gaps and unreliable measurements; a unique daily weather database at the city level was then obtained combining the 3 stations data. As a first step one of the three stations was chosen as the main source of data. More reliable pressure and solar radiation measurements recorded by each of the other two monitors were considered. Then daily averages were obtained by aritmetic mean (temperature, relative humidity, pressure),

geometric mean (wind speed, solar radiation), circual mean (wind direction), mode (wind direction - quadrants), maximum (wind speed), sum (rain). Finally missing daily values were imputed by averaging hourly data recorded 12h before and after the gap. Only rain levels were imputed as averages of those recorded at the other two stations.

After this initial exploratory stage of the analysis, spatio-temporal modeling of the multivariate normalized daily pollution data is performed within a Bayesian hierarchical framework proposed by Le and Zidek (2006), characterized by the use of time varying weather covariates and a semi-parametric spatial covariance structure. In the literature few examples of multivariate Bayesian models for space-time data are available. An interesting approach is proposed in Calder (2007), where a Bayesian dynamic factor process convolution model for multivariate spatio-temporal processes is described with an application to air quality data. This proposal results convenient when modeling highly dimensional air quality monitoring data. Key advantages of this framework are a descriptive parametrization of the cross-covariance structure of the space-time processes and the dimension reduction that allows full Bayesian inference to remain computationally tractable for large data sets. These features result from modeling the space-time multivariate data as realizations of linear combinations of underlying space-time fields. Another interesting approach to multivariate space-time data analysis is proposed in Shaddick and Wakefield (2002), where a hierarchical Bayesian model is described to obtain daily maps of four pollutants in the London area over the period 1994-1997. The authors use a dynamic linear modelling framework, caracterized by an exponential spatial covariance and a first order random walk nonstationary temporal structure. Both models admit any pattern of missing data, however they are characterized by a considerable computational complexity of the estimation procedure (MCMC and kernel convolution in Calder). In the following sections we point our attention to the model proposed by Le and Zidek (2006) as not only this is one of the few multivariate spatio-temporal statistical models for which several applications to air quality data are available (Le and Zidek 2006, Zidek et al. 2002), but it is also more

computationally efficient than the above mentioned approaches.

The paper is organized as follows. Missing data imputation is addressed in section 2. Section 3 explores the dependence on weather covariates and the temporal and spatial behaviour of the data. The modeling approach is briefly described in section 4 and in section 5 some results are reported. Section 6 is devoted to the discussion of the proposed strategy and to some concluding remarks.

2 Missing data treatment

Missing data is a ubiquitous problem in evaluating long-term experimental measurements, such as those associated with air quality monitoring. Spatio-temporal modeling often implies that such gaps in the measured data are filled or imputed. So far, no standardized method has been accepted and imputation methods used are largely dependent on the researchers' choice.

The objective of the method to be described in this section is to obtain a "full" database by imputing missing values. Here the basic idea is to preserve and exploit the spatial correlation of the observed concentrations, recursively estimating univariate daily spatial interpolation models for each pollutant, in order to predict missing data (Pollice and Jona Lasinio, 2008). This approach is taken to obtain an efficient tool for data pre-processing, reducing the computational complexity implied by considering a full spatio-temporal model. Alternatively the consideration of a unique marginal spatial model would lead to neglect the predictable changes in the spatial structure of the data along time. The procedure is based on Hierarchical Bayesian models models embracing properly defined spatial autocorrelation structures. These models can admit any pattern of missing measurements in a partially observed spatial process, as they provide predictive distributions that can be used for imputation. The usual linear mixed effects (LME) model, specified in two hierarchical levels, is chosen as the daily spatial interpolation model (Diggle and Ribeiro, 2007):

Level I - daily data process: Y is a p-dim Gaussian random field (GRF) representing the normalized daily mean concentrations of one pollutant

$$Y|\beta,\phi,\tau,\sigma^2 \sim N_p\left(\beta, V_y\left(\frac{\tau^2}{\sigma^2},\phi\right)\right)$$

Level II - prior specification:

- diffused priors for β and σ^2
- discrete priors on a specified reference grid for covariance structure parameters $\tau^2_{rel} = \tau^2/\sigma^2$ and ϕ

Due to the nonstandard prior structure, the predictive distribution has to be computed by numerical approximation: values of the covariance structure parameters τ^2 and ϕ simulated from their marginal discrete posterior distributions are plugged in the t-type predictive distribution obtained for the fully conjugate case.

The function `krige.bayes` in the `R` library `geoR` is used for the implementation of the following procedure making use of two daily spatial kinds of models specified as Bayesian LME's, namely *prediction models* and *estimation models*. Preliminarily, to properly set the prediction model priors for covariance structure parameters ϕ and τ^2_{rel}, a unique daily estimation model is fitted to available data and posterior estimates are obtained. Within a leave-one-out scheme daily spatial prediction models are then fitted and used to predict each missing observation. Priors are daily updated by posterior estimates obtained by the estimation model on the previous day. The spatial variation is thus believed to follow a sort of order 1 time dependence, with daily covariance parameter estimates depending stochastically on those of the day before. This recursive posterior-to-prior model estimation step is repeated updating missing observations until convergence is reached.

Letting y be the vector of normalized average daily concentrations for a specified pollutant on a certain day and J the set of indices denoting the missing monitoring stations, the whole procedure can be summarized in the following iterative algorithm:

step 0 A discrete uniform prior is chosen for τ^2_{rel} on the interval (0,1) with 0.1 increments, while ϕ is allowed to vary in a discrete sequence between 1 and 7 km with 0.5km incremental value and a reciprocal prior. For day 1 fit the estimation model to vector y where data corresponding to the missing stations are omitted. For days 2 to 365 fit the estimation model to vector y of the previous day, where data corresponding to the missing stations (z) are substituted. Obtain daily posterior estimates of ϕ and τ^2_{rel}.

step 1 For $i \in J$ let $y_{(i)}$ be obtained by omitting station i in the vector of daily observations y. Iteratively predict each y_i from $y_{(i)}$ using posterior estimates of ϕ and τ^2_{rel} obtained in the previous step for the prior specification of the prediction models. Store predicted values in vector z and substitute them to corresponding values in y.

step 2 Store the current z values in z_{old} and repeat step 1 to obtain a new z.

step 3 If $|z_{old} - z| < \varepsilon$ ($\varepsilon = 0.0001$) or the iterations number is ≥ 100 stop, otherwise repeat step 2 until convergence.

The entire procedure was investigated and compared to other approaches in Pollice and Jona Lasinio (2008). Fig. 2 shows a substantial agreement of the marginal distributions of the observed normalized daily average concetrations with those after missing data imputation for the Paolo VI site characterized by high NA rates (Tab. 2). This technique also shows a good capability towards spatial variation reconstruction and time dynamic preservation. Notice that the procedure we illustrated above was devised to include a calibration step which is not needed in the present case. Indeed in this case the algorithm is highly computationally efficient and convergence is reached within 1 or 2 iterations for almost all days in the study period.

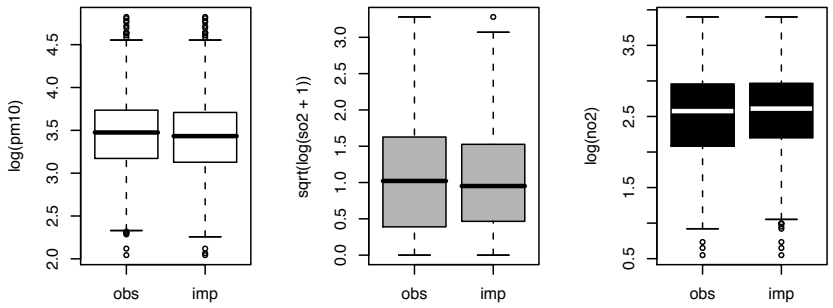


Figure 1: Boxplots of observed normalized pollutant concentrations before and after missing data imputation for the Paolo VI monitoring station. Similar results for the other 5 stations are available from the authors on request.

3 Exploring time and spatial patterns

In order to identify a suitable multivariate model structure to predict normalized daily concentrations we briefly investigate the relationship between meteorological covariates obtained as described in §1 and the pollutants concentrations.

The relevance of the covariates was verified fitting linear regression models: conditional OLS estimates were obtained for the normalized pollutant concentration levels at the 6 sites with weekday and month calendar variables and all the meteo covariates as explanatory variables. Concentration levels were overall significantly affected by the effects of weekday, calendar month, temperature, humidity, rain, maximum wind speed and wind direction quadrant (Tab. 3).

For the 3 pollutants a strong daily temporal dependence is expected. The time series are characterized by a strong daily time correlation structure, remarkably consistent across all sites. Autoregressive and other unpublished analyses lead to the adoption of a single AR(1) model for each pollutant, fitted across all 6 monitoring sites (Fig. 2). Residuals of a single AR(1) model for each pollutant, esti-

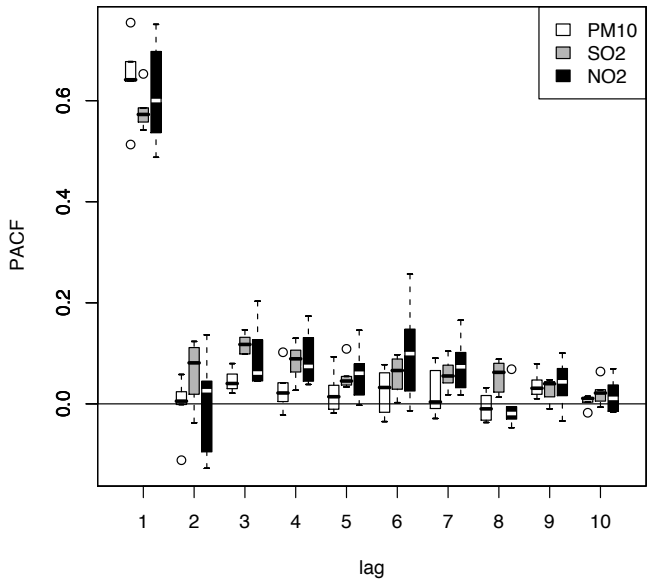


Figure 2: Boxplots of PACF's of spatially pooled normalized daily mean concentrations.

mated by pooling the 6 time series of normalized daily concentrations after imputation, don't show a significant correlation at lower lags.

The variation in the residuals can thus be expected to arise from variation due to space only. In order to separately model temporal and spatial variability we verify the absence of the so called *spatial correlation leakage* (Zidek et al., 2002) clearly shown in Fig. 3: the subtraction of the AR(1) temporal trend does not imply an overall decrease in the correlogram (the spatial structure is not diminished). Spatial modeling of the multivariate time-detrended pollutants concentration series is then dealt with in the next section.

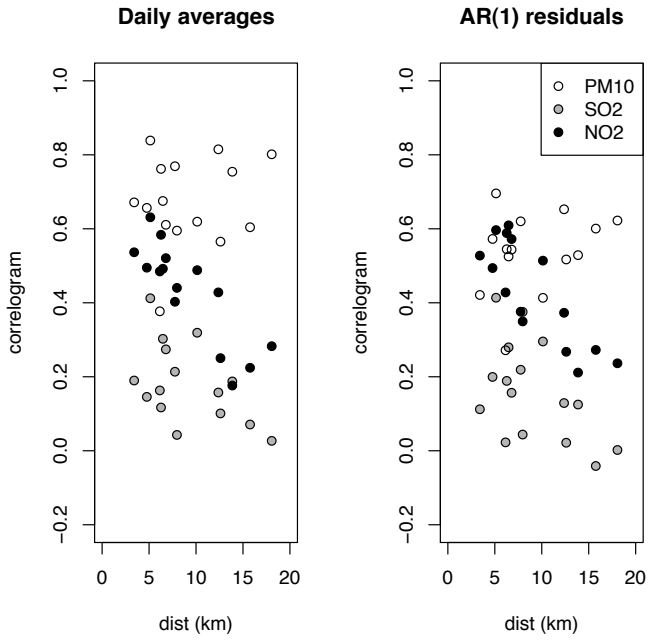


Figure 3: Correlograms for observed normalized daily average pollutant concentrations and residuals after the subtraction of the AR(1) temporal trend.

4 Spatial interpolation

The spatial predictive distribution was obtained by the multivariate Bayesian kriging-based model proposed by Le and Zidek (2006) and characterized by the use of time varying covariates and a semi-parametric nonstationary covariance structure. It is specified in the form of a Bayesian hierarchical model:

Level I - data process:

$$Y|Z,\beta,\Sigma \sim N_{spt}(Z\beta, I_t \otimes \Sigma)$$

Here Y is a spt-dimensional response vector containing the normalized daily mean concentrations of p pollutants at s gauged and ungauged sites for t time points. The $(spt \times spr)$-dimensional matrix $Z = I_{sp} \otimes \tilde{Z}$ contains sp replicates of common time-varying covariates $\tilde{Z}$ measured at one site (e.g. daily weather data at the city level), while regression coefficients in vector β are admitted to vary over sites. Σ is the between sites/pollutants covariance matrix, so that, due to the Kronecker structure, the responses in Y are assumed to be independent over time conditionally on the corresponding covariates in Z.

Level II - conjugate prior specification:

$$\beta|\beta_0,\Sigma,F \sim N_{rst}(\beta_0, F^{-1} \otimes \Sigma)$$

$$\Sigma|\Theta,\delta \sim \mathrm{IW}(\Theta,\delta)$$

where F^{-1} is the among covariates variance component of β and IW stands for the inverse Wishart distribution (can be substituted with its generalized version with multiple degrees of freedom parameters $\delta = (\delta_1,\ldots,\delta_k)$, representing uncertainty associated with k different operational periods).

Due to the previous conjugate specification, the explicit expression of the predictive distribution is obtained as a multivariate t-distribution

depending on hyperparameters β_0, F, Θ and δ. Such hyperparameters are estimated by the following two-step procedure:

step 1 At the gauged sites (monitoring stations) parameter estimates are obtained by EM marginal likelihood maximization (empirical Bayes/type-II MLE);

step 2 At the ungauged sites (grid points) the respective covariance and cross-covariance components of Σ are obtained by the Sampson-Guttorp method (Sampson & Guttorp, 1992). The method is based on constructing a thin-plate splines smooth mapping between locations in the geographic space, where stationarity of the multivariate random field is not assumed, to locations in a (virtual) new space where isotropy is assumed. Multidimensional scaling is used to obtain new locations for which the isotropy assumption is appropriate and an isotropic variogram model is fitted using the observed correlations and distances in the new space. The smooth mapping function, together with the isotropic variogram model enables to estimate the spatial dispersion between the stations and the ungauged sites.

The estimate of the multivariate spatial covariance is used to obtain the spatial predictive distribution. Its expectation or the mean of a specified number of simulations at selected grid points can be used to interpolate the daily fields.

The method has some clear theoretical advantages including the consideration of a very flexible spatial covariance structure and explicit expressions of posterior distributions enabling to avoid computationally cumbersome MCMC estimates. Computations are also made easy by a suite of R functions implementing the above estimation/prediction framework, available at `http://enviRo.stat.ubc.ca`. For the sake of completeness we report that the need for sparse missing data imputation and for filtering the time variability due to the conditional time-independence assumption increases the multi-step feature of the whole procedure with a consequent loss of control over its overall variability.

5 Some results

Missing data are imputed by the procedure in §2 and, in compliance with the conditional temporal independence assumption, the AR(1) time detrended daily residuals are used to estimate the model in §4 with weather and calendar covariates selected as in §3. Fig. 4 shows the result of applying the Sampson and Guttorp method to obtain estimates of the spatial covariances and cross-covariances: the deformation of the geographic space appears to be consistent with the presence of the sea in the south-western part of the study area (for details on the interpretation see Sampson and Guttorp, 1992). A compromise between the complexity of the mapping and the fit to the parametric variogram model in the isotropic virtual space leads to the choice of the amount of smoothing ($\lambda = 0.04$). On the other hand, the resulting predictions are not particularly sensitive to the choice of the smoothing parameter λ (our experience with several attempts and Sun et al, 1998).

The multivariate predictive distribution obtained by the estimated spatial covariance is used to interpolate the daily time-detrended normalized pollutants fields on a 400 points grid. These additional prediction locations belong to a 14×31 square lattice with 700m cell side, covering the whole area of interest. The predictive distribution is used to obtain expectations and 1000 simulations at each of the 400 grid-points on each of the 1095 days. The estimated AR(1) components of §3 for the three pollutants are then added to such interpolated residuals, completing the construction of the multivariate spatial predictor. Daily expectations and simulations summaries (means, standard errors, upper and lower quantiles, extremes) at each grid-point are considered as the final output of the modeling strategy and used to assess its behavior and to describe the spatio-temporal diffusion of the pollutants. According to the Bayesian posterior predictive p-values paradigm (Meng, 1994), daily credibility intervals are obtained by the percentiles of the 1000 simulations from the predictive distribution. Observed normalized daily concentrations fall outside the corresponding credibility intervals quite rarely, showing an over-

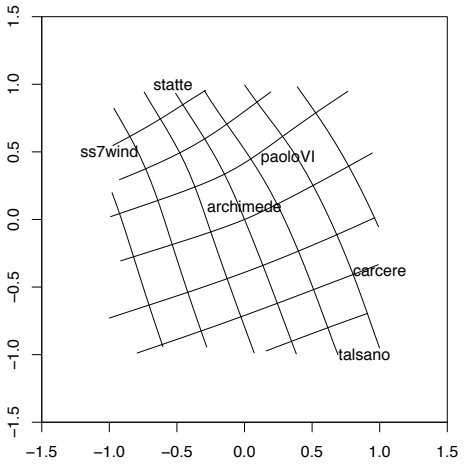


Figure 4: Biorthogonal grid characterizing the deformation of the geographic space obtained to reach approximate isotropy (normalized coordinates and $\lambda = 0.04$).

all compliance of the observed data with the simulations from the estimated predictive distribution (fig. 5).

A further assessment of the overall model fit is obtained comparing normalized pollutant concentrations at each monitoring station with predictions at the nearest grid-point by means of two model validation statistics (Carrol & Cressie, 1996):

$$CR_1 = S^{-1} \sum_s \frac{T^{-1} \sum_t \left(Y(s,t) - \hat{Y}(s,t) \right)}{T^{-1} \left(\sum_t \hat{\sigma}^2(s,t) \right)^{1/2}}$$

$$CR_2 = S^{-1} \sum_s \left(\frac{T^{-1} \sum_t \left(Y(s,t) - \hat{Y}(s,t) \right)^2}{T^{-1} \sum_t \hat{\sigma}^2(s,t)} \right)^{1/2}$$

When forecasts are accurate, CR_1 and CR_2 should be close to 0 and 1 respectively (Sahu & Mardia, 2005). The calculation of the first

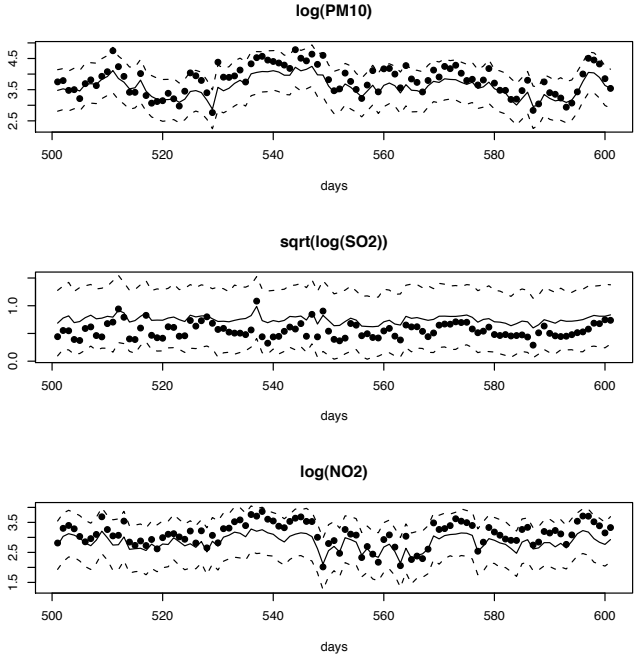


Figure 5: Normalized pollutant concentrations (black dots) for the Archimede monitoring station and those predicted at the nearest grid-point (solid line); dotted lines are 90% credibility intervals: days 500 to 600, i.e. 15/5/2006-23/8/2006.

order statistic CR_1 leads to very small values for the three pollutants (< e-04), providing some evidence for the marginal unbiasedness of the predictor. On the other hand the second order statistic CR_2 does not reach its optimal value, resulting equal to 0.56, 0.59 and 0.63 for PM10, SO2 and NO2 respectively. This shows that the variability of predictions underestimates the observed one, as it would be predictably implied by the intrinsic degree of smoothing of the spatio-temporal model.

Notwithstanding the evidence of some smoothing on the time scale (Fig. 5), the reconstruction of the time pattern obtained by the

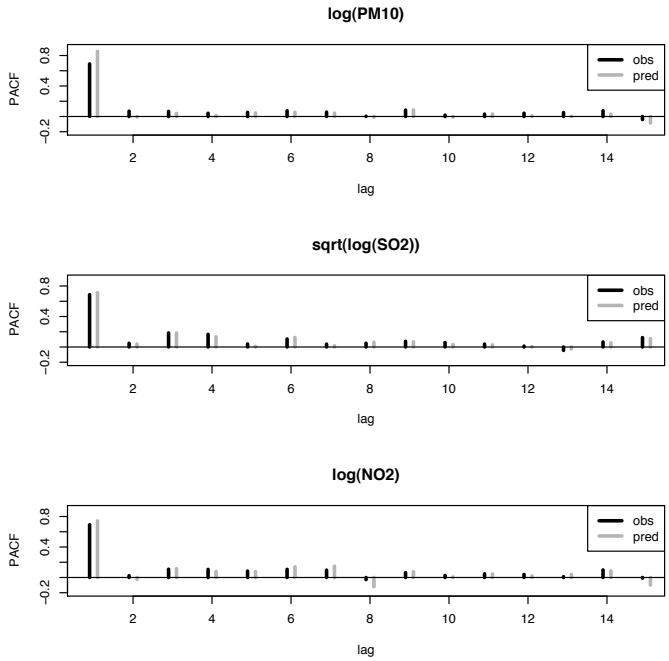


Figure 6: PACF's of normalized pollutant concentrations and of those predicted at the nearest grid-point for the Archimede monitoring station.

model is highly satisfactory: predictions look very accurate and the 90% credibility intervals look quite narrow and contain the observed values in the majority of days. Fig. 6 shows the substantial identity between the PACF's of the observed time series and those calculated for predictions at the nearest grid-point. Conditional results for all the other five monitoring stations (not shown) provide analogous evidence of this behaviour.

A first assessmet of the spatial behaviour of predictions was performed conditionally on the calendar day, by comparing observed concentrations at the six monitors with those predicted at the nearest grid-points. In Fig. 7 a substantial agreement between the two

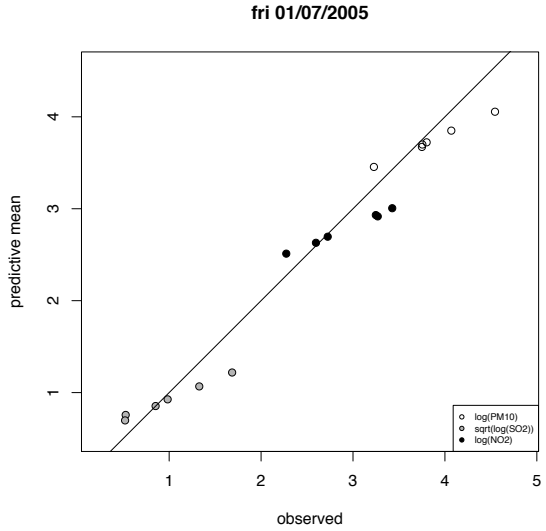


Figure 7: Normalized pollutant concentrations observed on 1-2/7/2005 and those predicted at the nearest grid-points.

series is shown, but some spatial smoothing is also evident.

Finally the concentration surfaces for the three normalized pollutants are obtained, by mapping multivariate model predictions for the 400 grid-points (Fig. 8). For all days spatial predictions show a strong connection with the wind direction. Most "hot-spots" are found in the vicinity of the iron plant (darker grey area in the maps) and the nearby Paolo VI monitoring station. Consistently with the empirical knowledge of the PM10 behavior, its peaks move south when the wind blows from the north-west direction and often lower concentrations are found when it rains considerably (more then 5mm). Maps of the simulations standard deviations and of the 90% credibility intervals (not shown), return daily evaluations of plausible values intervals and of the estimates quality.

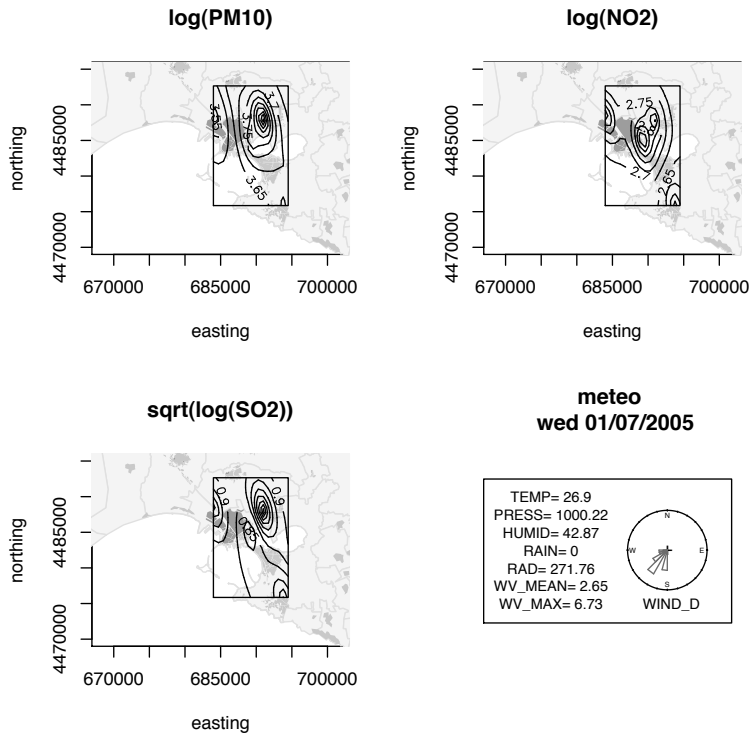


Figure 8: Expectation of the multivariate predictive distribution on a 14×31 square grid.

6 Discussion and concluding remarks

In this study daily estimates of three pollutants' concentration surfaces based on 6 monitoring stations were obtained in order to identify areas of higher concentration (hot spots), possibly related to specific anthropic activities.

Preliminary analyses involved addressing several data problems, mostly linked to the treatment of missing data and the selection of appropriate weather covariates. In section 2 we dealt with the first issue by a Bayesian kriging-based technique, using a hierarchical model

proposed by Diggle and Ribeiro (2007) enriched by a time-recursive prior structure (for a different approach see Fassò et al., 2007). Imputed values seem consistent with the experts empirical knowledge of the pollutants' behavior in the area.

Multivariate spatio-temporal modeling was then performed within a Bayesian hierarchical framework proposed by Le and Zidek (2006) and briefly described in sections 3 and 4. This approach is characterized by the use of time varying weather covariates and a semiparametric spatial covariance structure.

The proposed missing data treatment and the necessary removal of the temporal trend produce a composite estimation strategy for which it is particularly difficult to asses estimates precision. Indeed ignoring this aspect may seriously affect the final uncertainty evaluation and the use of an integrated model should be considered. On the other side the proposed approach is computationally efficient, unlike many more general Bayesian models involving complex MCMC simulation-based estimation procedures.

In general terms the proposed protocol returns coherent and satisfactory results with a reasonable computational effort.

References

Biggeri A, Bellini P, Terracini B. 2004. Metanalisi italiana degli studi sugli effetti a breve termine dell'inquinamento atmosferico. *Epidemiologia e Prevenzione*, **28**.

Calder CA. 2007. *Dynamic factor process convolution models for multivariate space-time data with application to air quality assessment.* Environmental and Ecological Statistics **14**: 229-247.

Carroll SS, Cressie N. 1996. *A comparison of geostatistical methodologies used to estimate snow water equivalent.* Water Resources Bulletin **32**: 267-278.

Diggle PJ, Ribeiro PJ. 2007. *Model-based Geostatistics*, Springer.

Fassò A, Cameletti M, Nicolis O. 2007. Air quality monitoring using heterogeneous networks. *Environmetrics* **18**: 245-264.

Le NZ, Zidek JV. 2006. *Statistical Analysis of Environmental Space-Time Processes*, Springer.

Meng XL. 1994. Posterior predictive p-values, *The Annals of Statistics* **22**, 3: 1142-1160.

Pollice A, Jona Lasinio G. 2008. Two approaches to imputation and adjustment of air quality data from a composite monitoring network. *GRASPA working paper* **30** (www.graspa.org).

Sahu SK, Mardia KV. 2005. A Bayesian Kriged Kalman model for short-term forecasting of air pollution levels. *Applied Statistics* **54**, 1: 223-244.

Sampson P, Guttorp P. 1992. Nonparametric estimation of non stationary spatial structure. *JASA* **87**: 108-119.

Shaddick G, Wakefield J. 2002. Modelling daily multivariate pollutant data at multiple sites. *Applied Statistics* **51**, 3: 351-372.

Zidek J, Sun L, Le N, Özkaynak H. 2002 Contending with space-time interaction in the spatial prediction of pollution: Vancouver's hourly ambient PM10 field. *Environmetrics* **13**: 595-613.

	PM10	SO2	NO2
PM10	$.53^2$	.02	.07
SO2	.08	$.39^2$	.04
NO2	.21	.18	$.62^2$

Table 1: Marginal correlation and covariance matrix for normalized pollutants (the variances lie on the diagonal, with covariances above and correlations below).

	Archimede	Carcere	PaoloVI	SS7wind	Statte	Talsano
PM10	321 (29)	98 (09)	143 (13)	183 (17)	199 (18)	20 (02)
SO2	183 (17)	109 (10)	176 (16)	206 (19)	93 (08)	25 (02)
NO2	209 (19)	120 (11)	202 (18)	214 (20)	159 (15)	71 (06)

Table 2: Missing daily averages (%).

Response		wkd	mon	tem	hum	rai	mwv	wdq
$log(\text{PM10})$	df	6	11	1	1	1	1	3
	F	1.04	14.26	214.35	17.07	123.96	61.23	11.99
	p	0.40	0.00	0.00	0.00	0.00	0.00	0.00
$\sqrt{log(\text{SO2}+1)}$	df	6	11	1	1	1	1	3
	F	0.75	10.70	35.82	221.34	2.18	50.23	28.10
	p	0.61	0.00	0.00	0.00	0.14	0.00	0.00
$log(\text{NO2})$	df	6	11	1	1	1	1	3
	F	2.80	15.80	7.02	0.26	35.45	30.66	34.06
	p	0.01	0.00	0.01	0.61	0.00	0.00	0.00

Table 3: ANOVA tables for the marginal linear regression models of normalized pollutant concentrations on calendar and meteo covariates (weekday, month, temperature, humidity, rain, maximum wind velocity and wind direction quadrant), concerning the Talsano monitoring station. Similar results for the other 5 stations are available from the authors on request.