Stat Methods Appl (2012) 21:75–91
DOI 10.1007/s10260-011-0178-8

Bayesian univariate space-time hierarchical model for mapping pollutant concentrations in the municipal area of Taranto

Serena Arima · Lorenza Cretarola · Giovanna Jona Lasinio · Alessio Pollice

Accepted: 22 September 2011 / Published online: 20 October 2011

Abstract An analysis of air quality data is provided for the municipal area of Taranto (Italy) characterized by high environmental risks as decreed by the Italian government in the 1990s. In the context of an agreement between Dipartimento di Scienze Statistiche—Università degli Studi di Bari and the local regional environmental protection agency air quality, data were provided concerning six monitoring stations and covering years from 2005 to 2007. In this paper we analyze the daily concentrations of three pollutants highly relevant in such an industrial area, namely SO_2, NO_2 and PM_{10}, with the aim of reconstructing daily pollutants concentration surfaces for the town area. Taking into account the large amount of sparse missing data and the non normality affecting pollutants' concentrations, we propose a full Bayesian separable space-time hierarchical model for each pollutant concentration series. The proposed model allows to embed missing data imputation and prediction of pollutant concentration. We critically discuss the results, highlighting advantages and disadvantages of the proposed methodology.

S. Arima (✉)
Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza,
Sapienza Università di Roma, Rome, Italy
e-mail: serena.arima@uniroma1.it

L. Cretarola
Presidenza del Consiglio,
Sapienza Università di Roma, Rome, Italy

G. Jona Lasinio
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Rome, Italy

A. Pollice
Dipartimento di Scienze Statistiche Carlo Cecchi, Università degli Studi di Bari "Aldo Moro",
Bari, Italy

Keywords Bayesian hierarchical model · Bayesian prediction · MCMC · Pollutant concentration · State-space model

1 Introduction

The analysis of the dynamics of airborne particulate mattern (PM_{10}) concentration, sulphur dioxide (SO_2) and nitrogen dioxide (NO_2) is a central issue in environmental monitoring. In fact, several epidemiological studies have shown that personal exposure to these pollutants has effects on lung functions, especially on children, the eldery and asthmatics.

Because of these dangerous effects on the health, several studies have been conducted all over the world in order to monitor pollutants levels, estimating space time surfaces over areas of interest and evaluating the effects of meteorological factors (e.g. temperature, relative humidity, wind velocity) on pollutants concentration (Asrari et al. 2005; Bush et al. 2001; Rajkumar and Chang 2000).

Similar studies have been conducted also in Italy, focusing on those areas characterized by high environmental risks (Primerano et al. 2006). In this paper, we focus on the municipal area of Taranto (southern Italy): this area is 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 other environmental impacting activities are more deeply integrated within the urban area and have to do with the presence of a large commercial harbor 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. 2001). In the context of an agreement between Dipartimento di Scienze Statistiche—Università degli Studi di Bari and ARPA Puglia (the local regional environmental protection agency), air quality data for the municipal area of the city of Taranto were provided, belonging to different monitoring networks pertaining to the regional and municipal government and counting up to 25 monitoring stations on the whole (Primerano et al. 2006).

Pollutants continuously monitored by the stations include sulphur dioxide (SO_2), nitrogen oxide (NO_X) and nitrogen dioxide (NO_2), carbon monoxide (CO), benzene, PM_{10} and ozone. This study is focused on PM_{10}, SO_2 and NO_2 concentrations. Only six monitoring stations were considered, those producing the longest time series for the three pollutants.

The main goal of this work is the proposal and critical analysis of a hierarchical Bayesian modeling framework for mapping, each one independently from the others, the concentration of three pollutants in the air shed of the city of Taranto. Extending the idea recently presented in Cocchi et al. (2007), we propose a hierarchical spatio-temporal modeling approach to describe and map daily mean concentrations of PM_{10}, SO_2 and NO_2. The model explicitly describes the spatial and temporal relationships within the data and those between pollutant concentrations and meteorological variables. This feature allows a better understanding of the pollutants diffusion and generating processes that is crucial to infer on their sources and effects on human health. The same data have been analyzed in Pollice and Jona Lasinio (2010): there the authors adopted a multi-step procedure based on the combination of a multivariate hierarchical spatio-temporal model within a Bayesian framework proposed by Le and Zidek (2006) and an external missing data imputation procedure based on spatial interpolation, the latter carried on in the Bayesian framework too. On the other hand, here we fully benefit of the Bayesian approach treating the missing data as unknown parameters and embed their estimation in the proposed model.

The paper is organized as follows. In Sect. 2 we introduce our data and some exploratory analyses are illustrated. In Sect. 3 the proposed model is defined and similarities and differences with other methods proposed in literature are highlighted. Results are discussed in Sect. 4 and model checking procedures are reported in Sect. 5. Some concluding remarks are given in Sect. 6.

2 Data description

The present study is focused on three pollutants: PM_{10}, NO_2 and SO_2. The data set contains validated data for years 2005–2007 (1 January 2005–31 December 2007), available for only 6 monitoring stations managed by the Apulia 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 the same time period and for 3 weather monitoring stations. Our main objective is to integrate pollution and meteorological data in order to summarize the spatial behavior of the pollution diffusion processes over the area of the municipality for the study period.

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 in order to reach approximate marginal Gaussianity: the square roots of the logs of SO_2 and the logs of PM_{10} and NO_2 daily average concentrations were considered. Original and transformed data are shown in Figs. 1, 2 and 3: the suggested transformations strongly improve the normality of the data.

Another interesting problem is the presence of a substantial percentage of missing data in the three pollutants. In Table 1 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). Missing data are embedded in the model as unknowns and estimated jointly with the other parameters. Available weather data are also 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 arithmetic mean

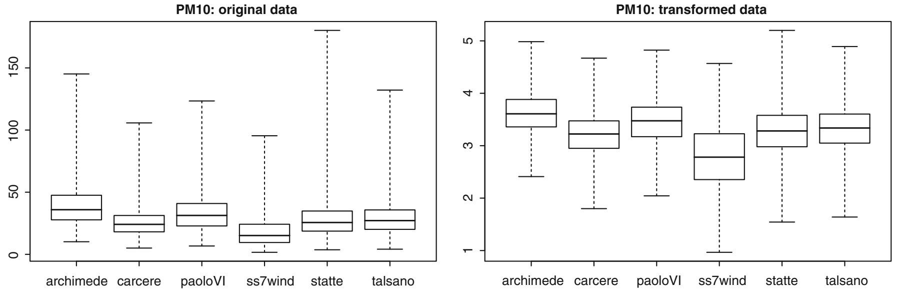


Fig. 1 PM_{10} concentrations for the six stations: original data (*left panel*) and transformed data (*right panel*). The simple logarithmic transformation strongly improves the normality of the data

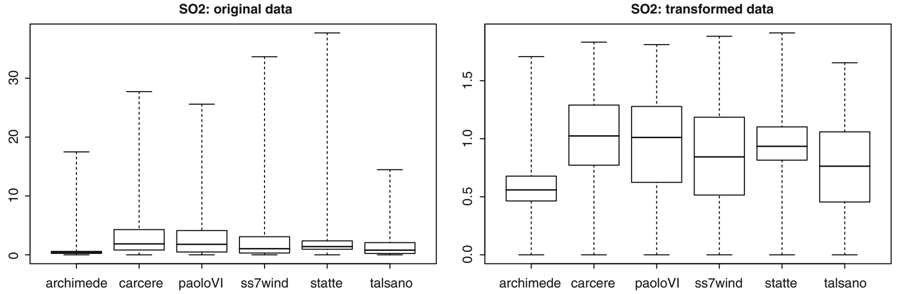


Fig. 2 SO_2 concentrations for the six stations: original data (*left panel*) and transformed data (*right panel*). The data were transformed by taking the square root of the logarithmic transformation; this transformation strongly improves the normality of the data

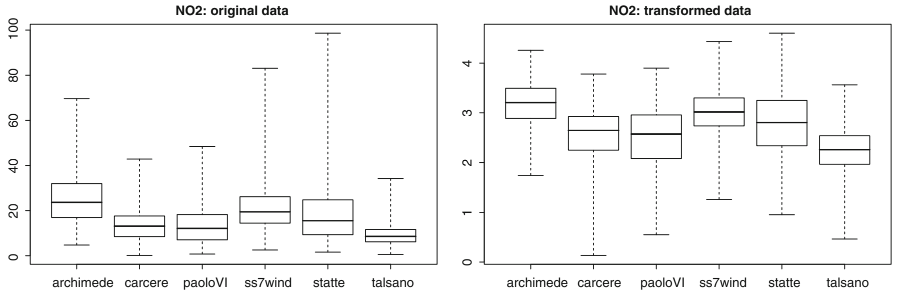


Fig. 3 NO_2 concentrations for the six stations: original data (*left panel*) and transformed data (*right panel*). The simple logarithmic transformation strongly improves the normality of the data

(temperature, relative humidity, pressure), geometric mean (wind speed, solar radiation), circular mean (wind direction), mode (wind direction—quadrants), maximum (wind speed), sum (rain). Missing daily values were imputed by averaging hourly data recorded 12 h before and after the gap. Only rain levels were imputed as averages of those recorded at the other two stations. Notice that differently from the pollutants, missing weather data were imputed due to their role in the proposed model.

Table 1 Missing daily averages (%)

	Archimede	Carcere	Paolo VI	SS7Wind	Statte	Talsano
PM_{10}	321 (29%)	98 (9%)	144 (13%)	184 (17%)	199 (18%)	23 (2%)
SO_2	183 (17%)	109 (10%)	176 (16%)	206 (19%)	93 (8%)	25 (2%)
NO_2	209 (19%)	120 (11%)	202 (18%)	214 (20%)	159 (14%)	71 (7%)

Pollutants are treated as response variables and weather data as covariates. In order to estimate missing covariates data a missing data mechanism has to be specified and other parameters have to be added. We decide to externally impute the missing data in the covariates in order to keep the model simpler.

Not all variables were considered like possible covariates for the construction of the models. Their relevance was verified by fitting linear regression models: conditional OLS estimates were obtained for the normalized pollutants concentrations at the 6 sites with weekday and month calendar variables and all meteorological 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. To these, we added the spatial coordinates of the sites.

3 The modelling approach

In this Section we sketch the basic structure of the proposed hierarchical model. Space-time hierarchical models have a relatively long history in the statistical literature. Starting from the seminal paper by Wikle et al. (1998), going to book chapters (Banerjee et al. 2004, ch. 8) and a variety of research articles often dealing with challenges arising from specific applied problems. Most of these papers start from the so called geostatistical approach where the observations are modeled as a partial realization of a spatio-temporal, typically Gaussian, random function

$$Z(s;t) = \mu(s,t) + e(s,t), \quad s \in \Re^d, \ t \in \Re$$

where $\mu(s,t)$ is the mean structure and $e(s;t)$ denotes the residuals, each elements of this formalization is then modeled to include physical knowledge, specific space-time dynamics and more empirical knowledge such as covariate effects deduced from available data. Z can be univariate or multivariate with the basic assumption is that second moments exist and are finite. Several simplifying hypothesis are made, as separability between space and time, stationarity etc. raising often many issues in terms of justifiability and model quality. Modeling efforts are,usually, directed to represent a large scale variation (trend, level in space and time) and the second moments (covariance structures), using a hierarchy of equations each dealing with one aspect of the phenomenon. These models are a direct mathematical representation of the reductionist paradigm. As each equation directly connects to a *part of the system*, it is important to stress that hierarchical models do not bound the system representation

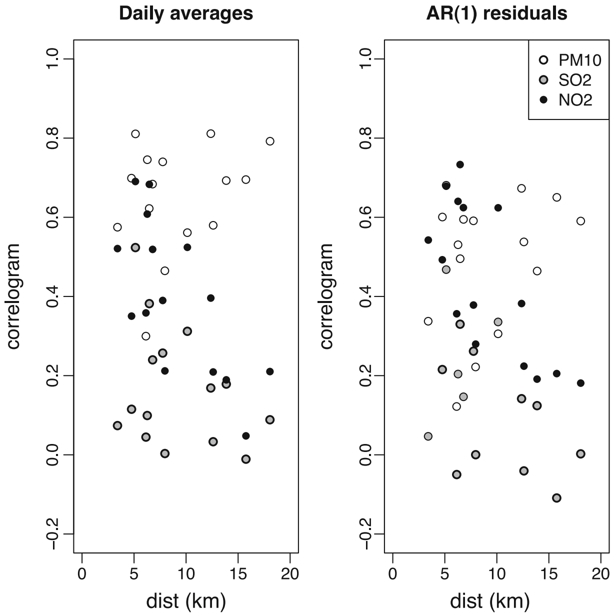


Fig. 4 Correlograms for observed normalized daily average pollutant concentrations and residuals after the subtraction of the AR(1) temporal trend

to a *sum of parts*: estimates and predictions uncertainty is obtained by hierarchically modeling the uncertainty of all t he parts of the system, allowing a wider perspective on the environmental-ecological object. In practical terms all these models share a common problem: computational complexity. No simple procedures are available and few ready-made software packages exist. This lack of software is justified by the necessary specificity of each case. In the Bayesian framework separable models can often be implemented in WinBugs (Shaddick and Wakefield 2002; Cocchi et al. 2007) allowing for a reasonable compromise between computational efficiency and easiness of implementation. We acknowledge that the separability between space and time is a strong assumption and this assumption is often not realistic from an empirical point of view [see for example Brown et al. (2000)]. However, when it applies, the model formulation is simplified and the computational complexity of the estimation procedure is reduced. In order to verify the separability assumption, we take advantage of the results in Pollice and Jona Lasinio (2010): the authors discussed the separability assumption using the so-called spatial correlation leakage proposed in Le and Zidek (2006). Figure 4 shows the correlograms for observed normalized daily average pollutant concentrations (left panel) and residuals after the subtraction of the AR(1) temporal trend for our data (right panel): since the subtraction of the AR(1) temporal trend does not imply an overall decrease in the correlogram, the spatial correlation leakage is absent and the separability between space and time assumption is supported by our data.

All these attractive features, and the mentioned missing data treatment typical of the Bayesian framework, lead us to our proposal.

3.1 The proposed model

In this section we define the proposed Bayesian hierarchical model specifying each level.

Level 1: *observed data model*
Suppose that a pollutant is observed at S spatial locations and T time points, along with a set of q meteorological variables. Let Y_{ts} denote the observed of one of the three pollutants on day t ($t = 1, \ldots, T$) at spatial location s ($s = 1,\ldots,S$); let (C_{1s}, C_{2s}) be the spatial coordinates of site s and let $\mathbf{X_t}$ be the q-dimensional vector of meteorological variables on day t.

At the first level of the hierarchy, conditional on the mean (μ_{ts}) and the measurement error variance (σ_s^2), observations are modelled as:

$$Y_{ts}|\mu_{ts}, \sigma_s^2 \sim N(\mu_{ts}, \sigma_s^2) \tag{1}$$

and

$$\mu_{ts} = \gamma_1 C_{1s} + \gamma_2 C_{2s} + \mathbf{X_t}'\beta + \theta_\mathbf{t} + \epsilon_\mathbf{ts} \tag{2}$$

Parameters γ_1 and γ_2 capture the large scale linear spatial trend, while vector β captures the dependence of pollutant levels on the covariates. θ_t represents a temporal random effect and the vector $\epsilon_{t.} = (\epsilon_{t1}, \epsilon_{t2}, \ldots, \epsilon_{tS})$ describes the spatial random effects at time t.

Level 2(a): *temporal model*
According to the results of some exploratory data analysis (Pollice and Jona Lasinio 2010), the time dynamic is represented as a first order autoregressive process:

$$\theta_t = \phi_1 \theta_{t-1} + \omega_t, \qquad \omega_t \sim N(0, \sigma_\theta^2) \tag{3}$$

Level 2(b): *spatial model*
We assume that the spatial and temporal processes are separable and that at each time t, the vector $\epsilon_{t.} = (\epsilon_{t1}, \epsilon_{t2}, \ldots, \epsilon_{tS})$ is a zero mean, isotropic Gaussian process with $S \times S$ correlation matrix Σ

$$\boldsymbol{\epsilon}_{\mathbf{t}.}|\boldsymbol{\sigma}_\epsilon^2, \boldsymbol{\Sigma} \sim \boldsymbol{MVN}(\mathbf{0}_S, \boldsymbol{\sigma}_\epsilon^2\boldsymbol{\Sigma}) \tag{4}$$

The sill parameter σ_ϵ^2 plays the role of the zero-distance variance. The *ss'* entry of the correlation matrix represents the correlation between sites s and s' and is assumed to be an exponential function of the distance $d_{ss'}$ between the two locations s and s':

$$\Sigma_{ss'} = \exp\left(-\phi d_{ss'}\right) \tag{5}$$

where $d_{ss'}$ is the distance between sites s and s'.

Level 3: *hyperpriors*
Model hierarchy is completed by prior specification for the hyperparameters. A Gaussian prior is assumed for the regression coefficients γ_1, γ_2 and β_i $(i = 1, \ldots, q)$. The exponential spatial structure in (4) is ruled by two parameters: the range ϕ and the sill $\sigma_\epsilon^2 = \frac{1}{\tau_\epsilon}$ of the covariance function. A flat truncated Normal prior distribution and a flat lognormal distribution have been chosen as prior distributions respectively for ϕ and τ_ϵ.

The ϕ_1 parameter was generated from a normal distribution with very small precision centered on the maximum likelihood estimate of a single AR(1) obtained by stacking the six monitoring stations recordings in a single series.

These three levels complete the proposed model. Our model can be considered as an extension of the model proposed in Cocchi et al. (2007) assuming a different time dynamic: we consider a first order autoregressive process rather than a simple random walk. This assumption has been supported by the data: as the posterior distribution of ϕ is clearly bounded away from 1 to 0 (it is concentrated around 0.73) showing some evidence of stationarity being well concentrated within the stationarity region.

Another important feature of the proposed approach is the missing data treatment: Pollice and Jona Lasinio (2010) analyzed the same data using a different multivariate model. However, they treated missing data by imputing them in a multi-step procedure. The obtained estimates and prediction resulted not fully satisfactory since the external imputation of missing data didn't allow to fully estimate uncertainty of the estimates. The parameters of the proposed model are estimated via Monte Carlo Markov Chain (MCMC) algorithm implemented in WinBUGS (Spiegelhalter et al. 1999).

4 Results

Starting from the model described in Sect. 3 we verified MCMC convergence for several model structures. Two separate chains of 50,000 iterations starting from overdispersed initial values were run for each model. A thinning interval of 25 and a burn-in period of 20,000 iterations were applied. Convergence was assessed by visual inspection of the chains sample trace plots, and by computing the Gelman and Rubin, Geweke and Raftery and Lewis statistics (Gilks et al. 1996).

The models point out the role of covariates in determining pollutants levels. Increases in the rain amount and maximum wind speed reduce PM_{10}, on the contrary temperature and relative humidity have positive coefficients, in accordance with the PM_{10} production process encouraged by high temperatures during warmer seasons. Also the wind direction is significantly related to the PM_{10} concentration, suggesting the presence of a transport phenomenon of particulate. Fitting the same model to NO_2 shows no significant influence of the rain amount and wind direction, while increases in temperature and relative humidity contribute to the pollutant's production, following a winter to summer reduction of its level. The model for SO_2 highlights how the rain amount and the calendar month have no influence on the pollutant level, while temperature, relative humidity and wind speed and direction contribute significantly. Spatial coordinates play the same role for PM_{10} and SO_2, showing a positive however, very small correlation with the pollutants level, suggesting the presence of a positive gradient in the NE direction complying with the possible effect of the sea breeze on the reduction of the levels of these two pollutants. While as far as NO_2 is concerned the spatial behavior is less clear, little variation can be found in the SW direction.

5 Model checking

Once the model has been estimated, the plausibility of the chosen model and the adequacy of model fitting have to be checked. Cross-validation is one of the most widely used tool to evaluate model fitting. However, in Bayesian framework cross-validation can be highly computationally expensive. In fact, in order to perform leave-one-out cross-validation, the model must be re-fit n times and fitting a Bayesian model even once can require long iterative computation. Hence, in a Bayesian context predictive checking is preferred. The Bayesian setting has the desirable feature of allowing us to discuss various critical aspects of the model from a predictive point of view: replicated data are generated from the predictive distribution once new parameter values are drawn from their posterior distributions and compared with the observed data.

We aim at evaluating the model's ability in predicting the average pollutants levels at unmonitored locations. In particular, daily normalized pollutant fields are interpolated on a 400 points grid: a 14 × 31 square lattice with 700 m cell side, covering the whole area of interest. Prediction of the pollutant concentration at site *s'* and time *t* is obtained by sampling from the posterior predictive distribution $p(\mu_{s't}|Y)$ whose components are:

$$\mu_{ts'}|Y = (\gamma_1|Y)C_{1s'} + (\gamma_2|Y)C_{2s'} + X_t^{'}(\beta|Y) + (\theta_t|Y) + (\epsilon_{ts'}|Y). \qquad (6)$$

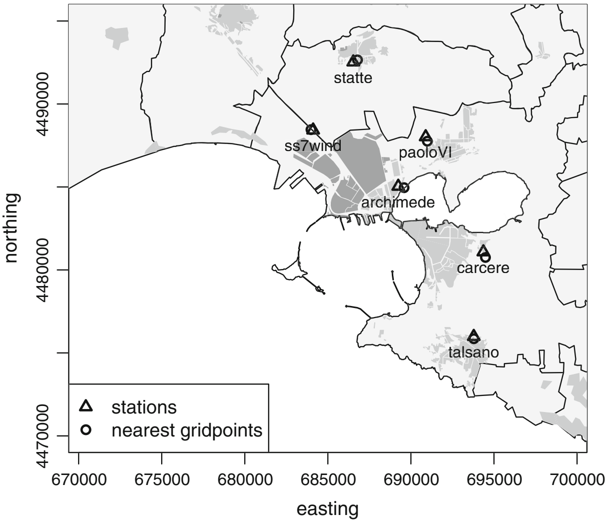


Fig. 5 Locations of the six monitoring stations and of the nearest grid points

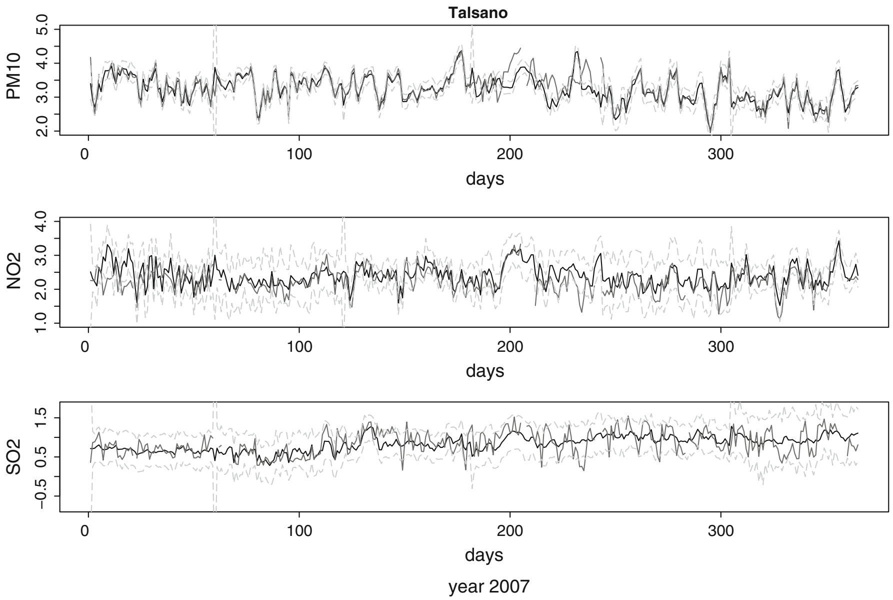


Fig. 6 Normalized observed pollutant concentrations (*dark grey line*) for the Talsano monitoring station and those predicted at the nearest grid point (*black line*); *dotted lines* are 95% credibility intervals. Year 2007

Table 2 RMSE's for the six monitoring stations over four time windows: SO_2

Years	Root mean squared error					
	Archimede	Carcere	Paolo VI	SS7Wind	Statte	Talsano
2005	0.4476	0.3801	0.5920	0.5955	0.3225	0.4905
2006	0.3427	0.3695	0.4570	0.4499	0.2428	0.3545
2007	0.3526	0.3062	0.4044	0.3555	0.1770	0.2570
2005–2007	0.3874	0.3557	0.4966	0.4844	0.2578	0.3813

Justification of the additive form of the predictive distribution is given in Shaddick and Wakefield (2002). Samples from the predictive distribution $p(\mu_{s't}|Y)$ are obtained via MCMC. We obtained 200 simulations at each of the 400 grid-points on each of the 1,095 days. Daily expectations and simulations summaries (means, standard errors, upper and lower 95% credibility interval limits) for the grid points closest to the six monitoring stations (see Fig. 5) are considered as the final output for the evaluation of the modeling strategy. In Fig. 6, the time dynamic of predicted and observed values together with the 95% credibility bounds are reported. The large majority of observed normalized daily concentrations fall inside the corresponding credibility intervals, showing an over-all compliance of the observed data with the simulations from the estimated predictive distribution.

In order to assess the model's ability in predicting pollutants levels we consider the following validation statistics:

Table 3 RMSE's for the six monitoring stations over four time windows: NO_2

Years	Root mean squared error					
	Archimede	Carcere	Paolo VI	SS7Wind	Statte	Talsano
2005	0.5881	0.5688	0.6638	0.5357	0.5226	0.5702
2006	0.6603	0.6554	0.6400	0.5422	0.5191	0.5437
2007	0.6326	0.4188	0.4558	0.5581	0.7772	0.3339
2005–2007	0.6276	0.5617	0.6021	0.5478	0.6039	0.4933

Table 4 RMSE's for the six monitoring stations over four time windows: PM_{10}

Years	Root mean squared error					
	Archimede	Carcere	Paolo VI	SS7Wind	Statte	Talsano
2005	0.5152	0.3722	0.2778	0.4510	0.2201	0.1179
2006	0.5180	0.2164	0.2038	0.6041	0.2499	0.1064
2007	0.4114	0.2853	0.3231	0.4884	0.2975	0.2495
2005–2007	0.4939	0.2997	0.2699	0.5172	0.2584	0.1862

– root mean squared error (RMSE)

$$RMSE_s = \sqrt{(MSE_s)} = \sqrt{\frac{\sum_{i=1}^{T}(Y_{ts} - \hat{Y}_{ts})^2}{T}},$$

where Y_{ts} represents the observed normalized pollutant concentrations at time t and monitoring location s and $\hat{Y}_{ts}$ represents predictions at time t and the nearest grid-point;

– CR_1 (Carroll and Cressie 1996)

$$CR_1 = \frac{1}{S}\sum_{s=1}^{S}\left\{\frac{1}{T}\sum_{t=1}^{T}\frac{(Y_{ts} - \hat{Y}_{ts})}{\hat{\sigma}_{ts}}\right\}$$

allows us to verify the unbiasedness of the predictors, it should be as close as possible to 0;

– CR_2 (Carroll and Cressie 1996)

$$CR_2 = \frac{1}{S}\sum_{s=1}^{S}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{(Y_{ts} - \hat{Y}_{ts})}{\hat{\sigma}_{ts}}\right)^2\right\}^{\frac{1}{2}}$$

verifies the accuracy of the mean squared prediction error and should be as close as possible to 1.

Tables 2, 3, 4 show RMSE values for the three pollutants, computed at the monitoring stations and over different time windows.

Table 5 CR_1 and CR_2 computed between monitoring stations and nearest grid points, over several time windows, for all pollutants

Year	Index	SO_2	NO_2	PM_{10}
2005	CR_1	−0.2350	−0.1385	0.0163
	CR_2	1.7630	1.3678	2.2205
2006	CR_1	−0.0003	−0.1474	−0.0931
	CR_2	1.6729	2.1789	2.3856
2007	CR_1	−0.0442	0.0909	0.6749
	CR_2	1.4732	2.3566	3.0347
2005–2007	CR_1	−0.0912	−0.0798	0.1885
	CR_2	1.6526	2.0071	2.6039

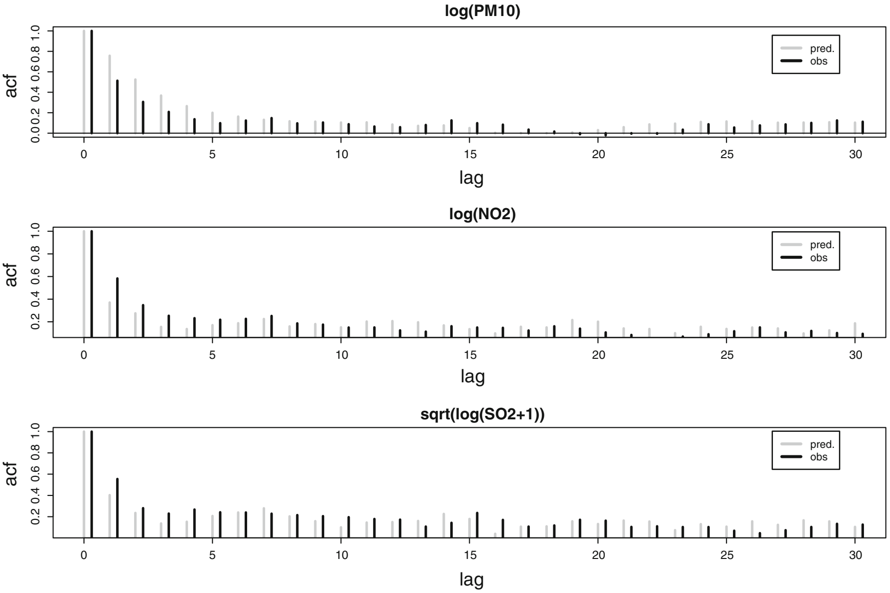


Fig. 7 ACF's of normalized pollutants and of those predicted at the nearest grid point for the Archimede monitoring station (*grey bars*)

The model predictions performance in terms of pollutants levels is satisfactory, in terms of pollutants normalized concentrations, their values are small for all sites and time windows, as RMSE's are expressed on the same scale as the normalized input data.

In Table 5 CR_1 and CR_2 values are reported. The best pollutants' level prediction is obtained for SO_2 in 200 and for NO_2 and PM_{10} in 2005. While an overall slight tendency to overestimation is shown when analyzing the three years at once for SO_2 and NO_2 ($CR_1 < 0$).

To analyze the model behavior with respect to the time dynamic, we examine the autocorrelation functions for observed concentrations and those predicted at the

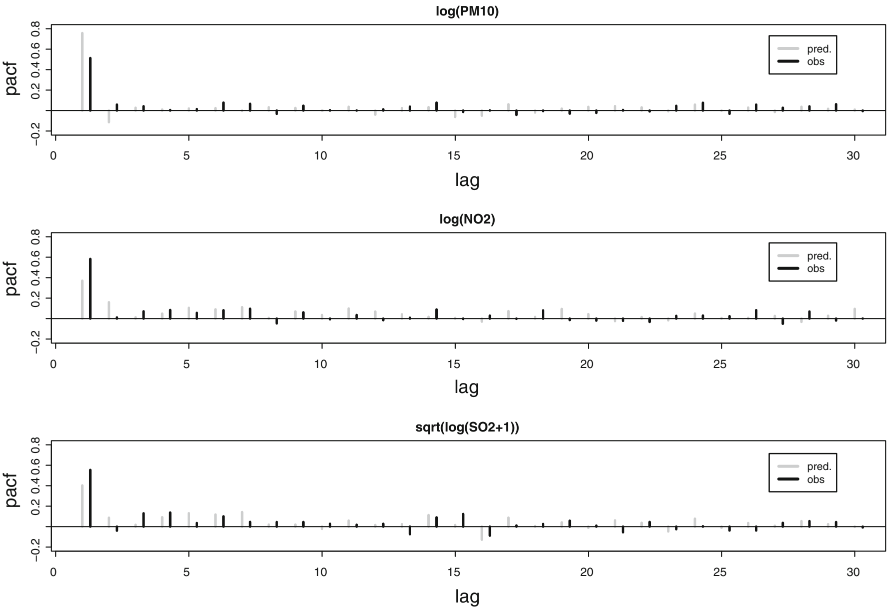


Fig. 8 PACF's of normalized pollutants and of those predicted at the nearest grid point for the Archimede monitoring station (*grey bars*)

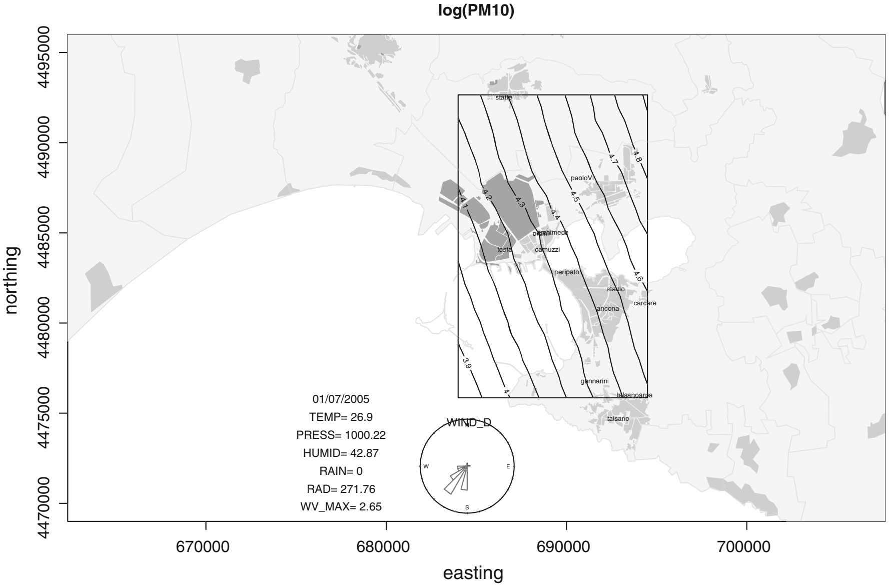
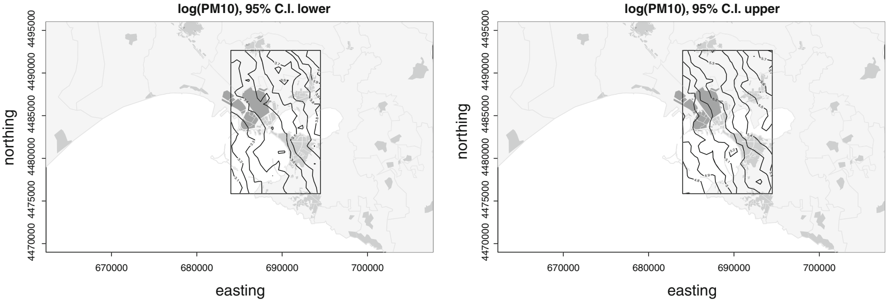


Fig. 9 Log PM_{10} July 1, 2005 map

nearest grid point. ACF's and PACF's are shown in Figs. 7 and 8, respectively for the Archimede monitoring station. Similar results are obtained with the other five monitoring stations. It must be noticed that observed ACF's and PACF's estimates are



lower 95% CI upper 95% CI

Fig. 10 Transformed PM_{10} July 1, 2005, 95% credibility interval maps

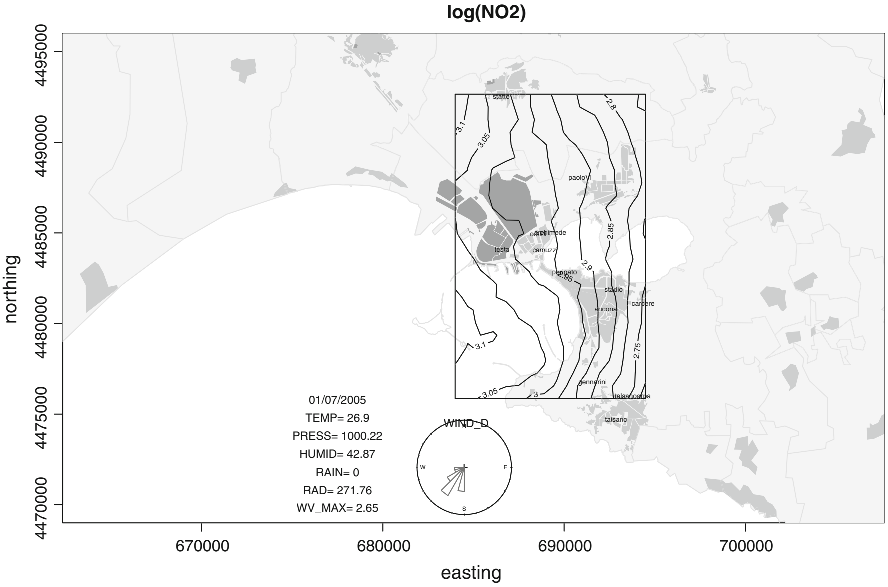
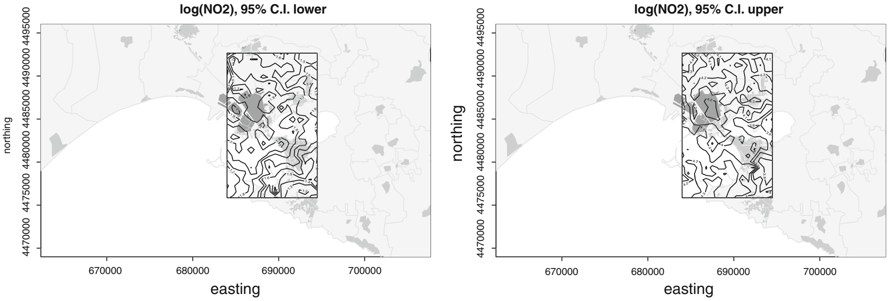


Fig. 11 Log NO_2 July 1, 2005 map



lower 95% CI uPPer 95% CI

Fig. 12 Log NO_2 July 1, 2005, 95% credibility interval maps

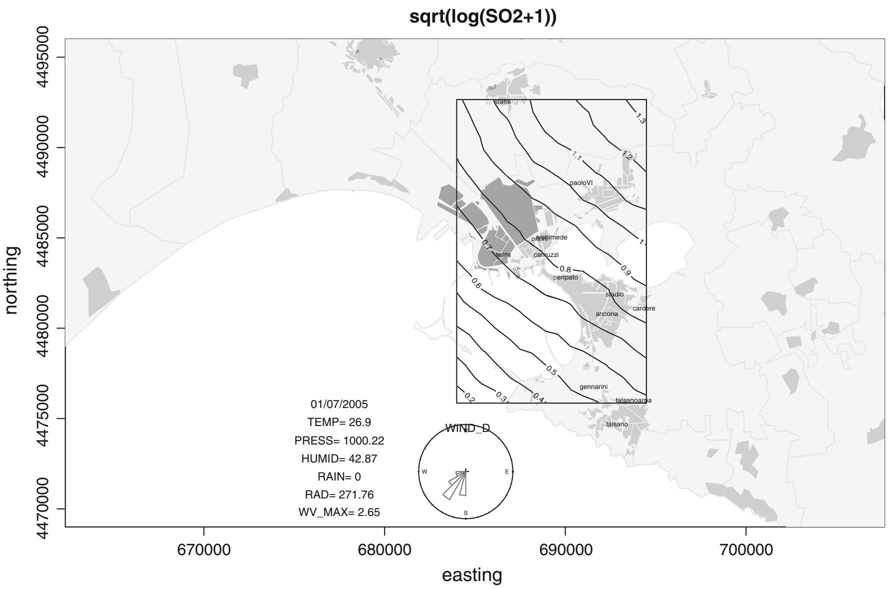
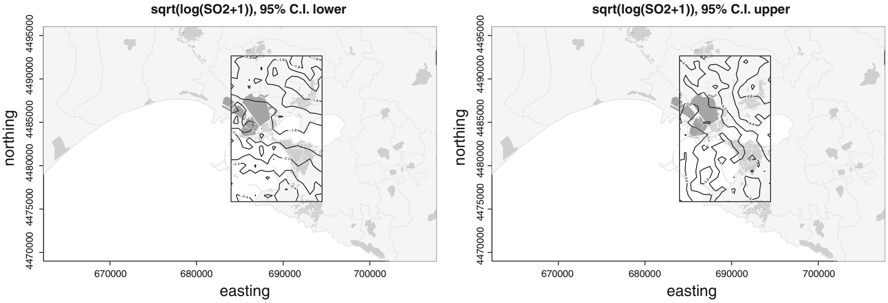


Fig. 13 Transformed SO_2 concentration December 16, 2007 map



lower 95% CI

uPPer 95% CI

Fig. 14 Transformed SO_2 July 1, 2005, 95% credibility interval maps

obtained from time series with a large number of missing data and can thus be unreliable. However, results are satisfactory on the whole. The autocorrelation structure of observed time series is well reproduced by all three models, while more discrepancies can be found in the partial autocorrelation.

As far as the spatial prediction is concerned we report some examples of maps for each pollutant and the corresponding 95% credibility intervals (Figs. 9, 10, 11, 12, 13, 14).

In the PM_{10} and SO_2 maps the influence of the wind direction appears clearly, while NO_2 flatter surface seems less influenced by this meteorological condition. This is in accordance with what was previously noticed concerning the significance of spatial coordinates and the influence of the sea breeze. All maps show very little spatial variation, however, they present highest pollutants levels where expected, according to the geomorphology and the locations of human activities over the area and to daily meteorological conditions. Examining credibility intervals for the three pollutants shows that while estimates of PM_{10} and NO_2 concentrations are acceptable, for the normalized SO_2 results are not completely satisfactory. The credibility intervals are wider then those obtained for the other two pollutants and their lower limit is sometimes negative, which is unrealistic. This maybe due to a lack of the model in capturing the large number of zeros recorded for SO_2.

6 Concluding remarks

In this work we analyze the behavior of a full Bayesian separable space-time hierarchical model adapted to predict normalized pollutant concentrations (PM_{10}, NO_2, SO_2) on a fine grid spanning the Taranto municipal area. The main advantages of the approach consist in its capability to easily handle missing data, properly reproduce the time dynamic and capture spatial information from the data. From the physical point of view the predicted maps have acceptable interpretation: all maps show very little spatial variation, however, they present highest pollutants levels where expected, according to the geomorphology and the locations of human activities over the area and to daily meteorological conditions. Larger values are detected near the main pollution source (dark grey area in the maps) and a decreasing gradient is detected following the wind direction. This is in accordance with typical diffusion of pollutants in air. However, we must acknowledge that results are not completely satisfactory for SO_2. Furthermore, results obtained for the three univariate models are not really comparable with those in Pollice and Jona Lasinio (2010), where a multivariate approach is considered. The authors are actually working on a multivariate version of the proposed model, that requires a considerable computational effort to be estimated: however, we believe that the lack of fit detected in the univariate model can be solved with a more comprehensive multivariate approach.

Acknowledgments The authors would like to thank ARPA Puglia for motivating this work and for providing the pollutants concentration and meteorological data. We are also grateful to the referee whose suggestions have improved the manuscript.

References

Asrari E, Ghole VS, Sen PN (2005) Study on the status of SO in the Tehran-Iran. J Appl Sci Environ Manag 10(2):75–82

Banerjee S, Carlin BP, Gelfand A (2004) Hierarchical modeling and analysis for spatial data. Monographs on statistics & applied probability. Chapman & Hall/CRC, New York

Biggeri A, Bellini P, Terracini B (2001) Metanalisi italiana degli studi sugli effetti a breve termine dell'inquinamento atmosferico. Epidemiologia e Prevenzione 28:1–72

Brown PE, Karesen KF, Roberts GO, Tonellato S (2000) Blur-generated non-separable space-time models. J R Stat Soc Series B 62:847–860

Bush T, Smith S, Stevenson K, Moorcroft S (2001) Validation of nitrogen dioxide diffusion tube methodology in the UK. Atmos Environ 35:289–296

Carroll SS, Cressie N (1996) A comparison of geostatistical methodologies used to estimate snow water equivalent. Wat Resour Bull 32:267–278

Cocchi D, Greco F, Trivisano C (2007) Hierarchical space-time modelling of PM_{10} pollution. Atmos Environ 41:532–542

Gilks WR, Richardson S, Spiegelhalter DJ (1996) Markov chain Monte Carlo in practice. Chapman & Hall, London 116–118

Le ND, Zidek JV (2006) Statistical analysis of environmental space-time processes. Springer, Berlin

Pollice A, Jona Lasinio G (2010) A multivariate approach to the analysis of air quality in a high environmental risk area. Environmetrics 21:741–754

Primerano R, Menegotto M, Di Natale G, Giua R, Notarnicola M, Assennato G, Liberti L (2006) Episodi acuti di inquinamento da PM10 nell'area ad elevata concentrazione industriale di Taranto. poster presented at Secondo Convegno Nazionale sul Particolato Atmosferico—PM2006, Florence, 10–13 Sept 2006

Rajkumar WS, Chang AS (2000) Suspended particulate matter concentrations along the East-West Corridor, Trinidad, West Indies. Atmos Environ 34:1181–1187

Shaddick G, Wakefield J (2002) Modelling daily multivariate pollutant data at multiple sites. Appl Statist 51(part 3):351–372

Spiegelhalter DJ, Thomas A, Best N (1999). WinBUGS Version 1.2 User Manual. MRC biostatistics unit. Software available at http://www.mrcbsu.cam.ac.uk/bugs/winbugs/contents.shtml

Wikle CK, Berliner LM, Cressie N (1998) Hierarchical Bayesian space-time models. Environ Ecol Stat 5:117–154