Bayesian spatio–temporal modeling of urban air pollution dynamics

Simone Del Sarto[1,*], M. Giovanna Ranalli[2], K. Shuvo Bakar[3], David Cappelletti[4], Beatrice Moroni[4], Stefano Crocchianti[4], Silvia Castellini[4]

[1] *Department of Economics, University of Perugia; delsarto@stat.unipg.it*
[2] *Department of Political Sciences, University of Perugia; giovanna.ranalli@stat.unipg.it*
[3] *CSIRO, Canberra, Australia; shuvo.bakar@csiro.au*
[4] *Department of Chemistry, Biology and Biotechnologies, University of Perugia; david.cappelletti@unipg.it, croc@impact.dyn.unipg.it*
[*] *Corresponding author*

Abstract. *This work deals with the spatio-temporal analysis of urban air pollution dynamics from the town of Perugia, Italy, using high-frequency and size resolved data on particular matter. Hierarchical Bayesian models are used that allow for an autoregressive term in time. Some preliminary results show that there is a significant spatio–temporal structure with a large first–order temporal correlation coefficient. Future analysis will concern the use of higher–order temporal auto–correlation structures and the introduction of the effect of some covariates.*

Keywords. *Hierarchical model; Air quality; Autoregressive structure; Gibbs sampling; Sensor data.*

1 Introduction

Urban pollution has an important impact on human health and environment. Investigating the behavior of pollutants and understanding air quality of particular geographical areas has been one of the central issues in environmental public policy and decision making. Air pollution often shows a spatio-temporal structure. Hierarchical Bayesian modeling provides useful tools to investigate spatial and/or temporal patterns also in large datasets [3, 4, 5]. There exist various types of spatial data, including spatio-temporal point referenced data [2], where observations are collected over time at several spatial locations, which vary continuously over a study area. In this paper we analyze data from the PMetro project (`http://www.pmetro.it`), which studies urban pollution dynamics in the town of Perugia (Italy) since September, 2012, using the `spTimer` R package [1]. Unlike classical monitoring of pollutants concentration using fixed stations, fast measure of gases and size resolved particulate matter (PM) is coupled with information on the evolution of urban microclimate and vehicular traffic fluxes. In particular, data is collected using an instrument located on a cabin of the Minimetro, a public conveyance that moves on

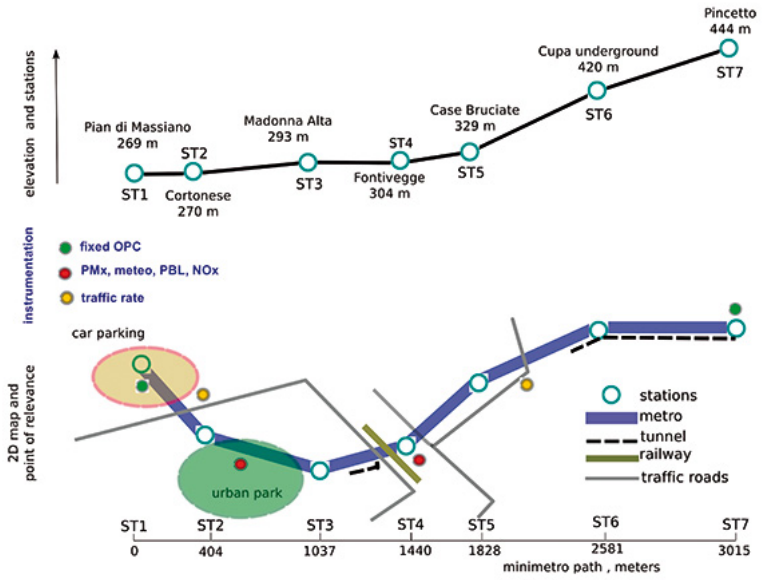

Figure 1: Schematic map of the minimetro path and of the sources of data. In the upper panel station names and elevation (meters a.s.l.) are indicated. In the lower panel a 2D sketch of the area suggests the main intersections with traffic roads and indicates also some points of interest (car parking, urban park, tunnels). The metro path is shown at the bottom together with distances along the path (in meters). Position and typology of the instruments employed in the project are also shown.

a monorail throughout the town. The paper is organized as follows. Section 2 illustrates the data analyzed in more detail and provides the structure of the spatio-temporal model employed. Then, Section 3 provides some preliminaries results obtained and directions for future work.

2 The data and the model

An OPC (Optical Particle Counter) integrated on a cabin of Minimetro is used to get a snapshot of the urban pollution dynamics along a sector of the town at high spatial and temporal resolution. Figure 1 provides all the detail on the metro path. It is about 3 km long with seven stations: a single travel takes about twenty minutes, so that each cabin runs along the same path about forty times a day. The path is outdoors for the most part and passes through parks, highly-traffic roads, residential areas and two tunnels. The OPC takes a sample of air every 6 seconds while the cabin moves on the monorail and counts the number of particles between 0.28 and 10 micrometers (μm, 10^{-6} meters) in 22 size channels. The position and speed of the cabin is continuosly recorded by the central control software of the Minimetro transport system and transmitted to the OPC data logger. Therefore, the sampling points are variable along the path and depend on the speed of the cabin, which is not constant along the path or during the day. In this paper we focus on the number concentration of *fine* particles, i.e. those with a diameter between 0.28 and 1.10 μm.

We analyze data collected on January and February, 2014: due to the different operation-time of the Minimetro on the basis of the day of the week, we cut the entire dataset to have observations from 6 am to 7 pm and we average the concentration measure in each hour, so that we have 14 hourly observations

per day. Furthermore, we divide the entire path of the Minimetro into $n = 45$ equally-spaced spatial segments, for which we determine the coordinates in Latitude and Longitude. Finally, we have measurements available for 23 days for each month, each day has 14 hourly observations, each hour has 45 observations relative to spatial segments: the final dataset is made up of 28,980 observations (46 days $\times$ 14 hours $\times$ 45 segments). It is possible to have some missing data due to maintenance and/or malfunctioning of the OPC, but we can consider missingness to be completely at random.

Let l and t denote the two units of time, where $l = 1,\ldots,46$ denotes the longer unit, i.e. the day, and $t = 1,\ldots,14$ denotes the shorter unit, i.e. the hour. Let $Z_l(\boldsymbol{s}_i,t)$ be the observed point referenced data and $O_l(\boldsymbol{s}_i,t)$ be the true value corresponding to $Z_l(\boldsymbol{s}_i,t)$ at space segment $\boldsymbol{s}_i$, $i = 1,\ldots,n = 45$ at time denoted by the two indices l and t. Then let $\boldsymbol{Z}_{lt} = (Z_l(\boldsymbol{s}_1,t),\ldots,Z_l(\boldsymbol{s}_n,t))^T$ and let $\boldsymbol{O}_{lt}$ be defined similarly in terms of $O_l(\boldsymbol{s}_i,t)$. For our analysis, we use the hierarchical autoregressive model proposed in [6] whose specification is as follows:

$$\boldsymbol{Z}_{lt} = \boldsymbol{O}_{lt} + \boldsymbol{\epsilon}_{lt}, \tag{1}$$

$$\boldsymbol{O}_{lt} = \boldsymbol{\beta}_0 + \rho \boldsymbol{O}_{lt-1} + \boldsymbol{\eta}_{lt}, \quad l = 1,\ldots,46 \quad t = 1,\ldots,14, \tag{2}$$

where $\boldsymbol{\epsilon}_{lt}$ is the so called nugget effect (or the pure error term) and is assumed to be independently normally distributed $N(\boldsymbol{0},\sigma^2_\varepsilon \boldsymbol{I}_n)$, where σ^2_ε is the unknown variance and $\boldsymbol{I}_n$ is an identity matrix of order n; $\boldsymbol{\beta}_0$ is a common intercept term. Furthermore, we also assume that the spatio-temporal random effects, $\boldsymbol{\eta}_{lt}$, follow a normal distribution $N(\boldsymbol{0},\boldsymbol{\Sigma}_\eta)$ independently in time, where $\boldsymbol{\Sigma}_\eta = \sigma^2_\eta \boldsymbol{S}_\eta$, with σ^2_η is the site invariant spatial variance and $\boldsymbol{S}_\eta$ is the spatial correlation matrix. This matrix can be determined using the general Matérn correlation function, or more simply it can have exponential form, depending on a parameter ϕ, which determines the rate of decay of the correlation as the distance between two locations increases [2]. Finally, ρ denotes the unknown temporal correlation parameter which lies in the interval (-1, 1).

3 Preliminary results and future developments

To estimate model (1)-(2) above, we use the R package `spTimer` [1] version 1.0-1 using a Gibbs sampling approach, with default values for the starting values for the model parameters and for the hyperparameters values of the prior distributions (for more details on this see [1]). Matrix $\boldsymbol{S}_\eta$ is determined using the exponential correlation function. For the spatial decay parameter ϕ we use a plug-in value equal to $-\log(0.05)/d_{\max}$, where $d_{\max}$ is maximum distance between locations. Given the skew distribution of the number of fine particles, we use a logarithmic transformation for the response variable. Furthermore, MCMC is run for 10,000 iterations with burn-in 1,000. The traceplots for the model parameters are reported in Figure 2 together with posterior density estimates.

The chains provide evidence of a good convergence of the model for all parameters. The posterior density estimates show that all parameters are statistically significant. We can also observe a strong spatio-temporal effect with the spatial variance σ^2_η having mean 2.407, significantly higher that the nugget effect σ^2_ε (mean equal to 0.059). Finally, there is also evidence of a strong positive autoregressive term in time, with ρ having posterior mean equal to 0.827.

These are just first results that are the basics for further analysis. First we wish to conduct a sensitivity analysis to better understand the role and provide a good plug-in value for the decay parameter ϕ. Then, since we have covariates, we wish to change equation (2) to

$$\boldsymbol{O}_{lt} = \rho \boldsymbol{O}_{lt-1} + \boldsymbol{X}_{lt}\boldsymbol{\beta} + \boldsymbol{\eta}_{lt}, \quad l = 1,\ldots,46 \quad t = 1,\ldots,14,$$

where $\boldsymbol{X}_{lt}$ is a matrix with n rows and, say, p columns (the number of covariates) and $\boldsymbol{\beta}$ is the p-vector of the regression coefficients. This will enable us to understand the effect of some covariates, such as

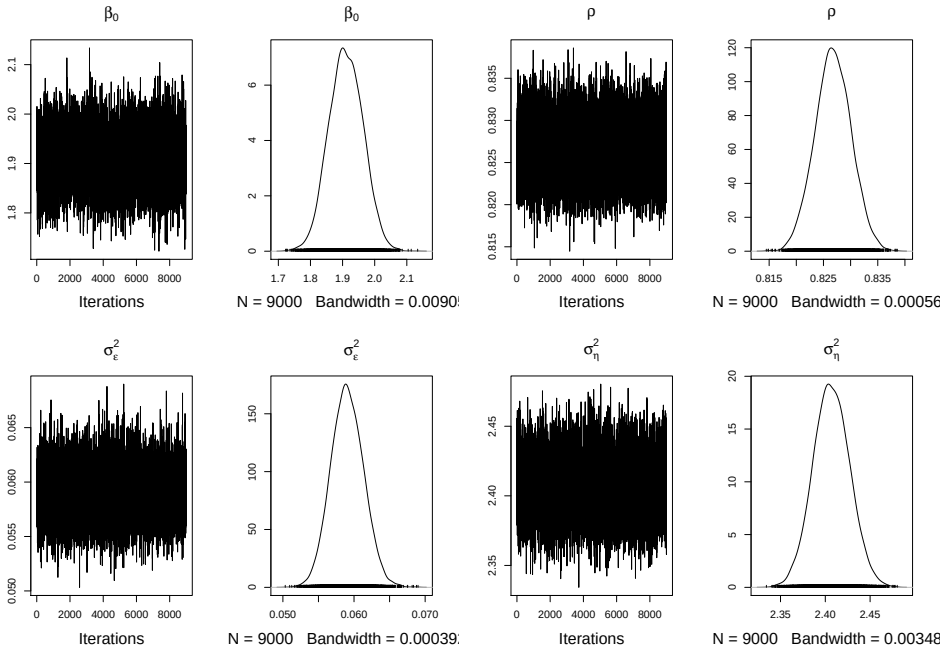

Figure 2: MCMC traceplots and density estimates for model parameters

the temperature or the relative humidity measured with instruments placed on the Minimetro cabin, or the altitude or the typology of the spatial segment. Moreover, these first results provide evidence of a significant first-order temporal autocorrelation so that we would like to introduce higher order autocorrelation terms. Once a the relatively best model has been picked, this approach allows for producing maps of concentrations over the urban area covered by the Metro path, and also for predicting maps for future time points. Finally, since the cabin also measures NO concentration at the same space-time resolution, it would be of great interest, in addition, to model it jointly with the fine particle concentration using a multivariate space-temporal model.

References

[1] Bakar, K.S. and Sahu, S.K. (2014). spTimer: Spatio-Temporal Bayesian Modelling Using R. *Journal of Statistical Software*, in press.

[2] Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton.

[3] Banerjee, S., Gelfand, A.E., Finley, A.O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society B* **70**, 825–848.

[4] Sahu, S.K. and Bakar, K.S. (2012a). A Comparison of Bayesian Models for Daily Ozone Concentration Levels. *Statistical Methodology* **9**(1), 144–157.

[5] Sahu, S.K. and Bakar, K.S. (2012b). Hierarchical Bayesian Autoregressive Models for Large Space Time Data with Applications to Ozone Concentration Modelling. *Applied Stochastic Models in Business and Industry* **28**, 395–415.

[6] Sahu, S.K., Gelfand A.E. and Holland D.M. (2007). High-Resolution Space-Time Ozone Modeling for Assessing Trends. *Journal of the American Statistical Association* **102**, 1221–1234.