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.

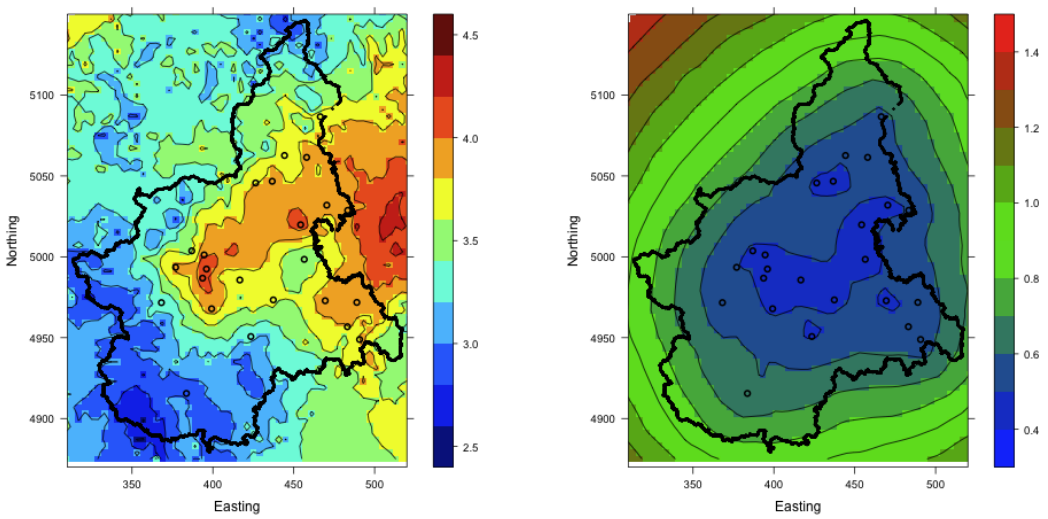

Figure 1: Map of the PM_{10} posterior mean on the logarithmic scale (left) and standard deviation (right) for January 29th, 2006.

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.