A spatio-temporal model for air quality mapping using uncertain covariates [1]

Michela Cameletti
Università degli Studi di Bergamo

Stefania Ghigo, Rosaria Ignaccolo
Università degli Studi di Torino, ghigo@econ.unito.it

Abstract: Particulate matter (PM) is one of the most critical air pollutants because of its effects on the human health and the environment. It is well known that covariates, such as meteorological and geographical variables, have a significative influence on PM concentration. In this work we model PM concentration, measured by the monitoring network in Piemonte, taking into account the uncertainty of covariates that are output of a deterministic model chain, by means of a spatio-temporal error-in-variables model. The aim is to map the PM concentration random field all over Piemonte region considering all the uncertainty sources, i.e. the error related to the PM measurements and the covariate simulation as well as the error coming from the spatial prediction procedure.

Keywords: Error-in-variables model, Bayesian hierarchical model, MCMC

1 Introduction and motivating case study

The aim of this paper is to provide a spatio-temporal model of PM concentration, observed by a monitoring network, as function of some significative covariates (such as meteorological variables) given as output of a deterministic modeling system. While it is routine to consider that PM measurements are subject to an instrumental error, it is not usual to take into account the uncertainty of numerical model outputs. Usually such outputs are considered deterministic, thus known without error. However, numerical models try to reproduce reality but are affected by uncertainty related to initial conditions, parameters in model equations as well as model structure (Bayarri et al., 2009). To take into account these uncertainty sources we propose a spatio-temporal error-in-variables model (also known as measurement error model) where latent processes are introduced for modeling both the "true" PM and covariate fields. Our proposal is an extension of the models proposed in Van de Kassteele et al. (2006a, 2006b), where purely spatial error-in-variables models are considered in order to "correct" the numerical model outputs for nitrogen dioxide and particulate matter, respectively. Thus Van de Kassteele et al. (2006a, b) quantify the uncertainty of numerical model outputs, taking them as covariates in

[1] Work partially supported by Regione Piemonte.

a spatial model for the same pollutant. Instead, we want to take into account the uncertainty of exogenous covariates in air pollutant modelling.

In our case study, we consider daily particulate matter with an aerodynamic diameter of less than 10 μm (PM_{10}) measured at $n = 24$ sites and $T = 93$ days (from November 15, 2005 to February 15, 2006) in the Northern Italian region Piemonte. Moreover, we select $m = 10$ sites for validation purposes (see blue dots in Figure 1(a)). Because of the complex orography of the region, the pollutant dispersion is strongly affected by meteorological and geographical conditions. To take into account this relationship, we consider the following significative covariates (selected through a preliminary regression analysis): altitude (in m), coordinates (UTM, in km), daily mean wind speed (in m/s), daily mean temperature (in ^oK) daily maximum mixing height (in m) and daily emission rates of primary aerosols (in g/s). The time-varying covariates are simulated on a $4\,km \times 4\,km$ regular grid by a numerical model implemented by the environmental agency ARPA Piemonte (Bande et al., 2007) and are available at the monitoring sites as well. These numerical output covariates are introduced in our model with errors, whereas the constant in time covariates are supposed to be known without error.

2 The error-in-variables model

Let $y(s_i, t)$ and $x_k(s_i, t)$ denote, respectively, the measured PM_{10} concentration and the simulated value of the $k-$th covariate at location s_i and time t, with $i = 1, \ldots, n$, $t = 1, \ldots, T$ and $k = 1, \ldots, K$. Assuming that both $y(s_i, t)$ and $x_k(s_i, t)$ are affected by an additive error, we define the following equations

$$y(s_i, t) = \eta(s_i, t) + \varepsilon_y(s_i, t) \tag{1}$$

$$x_k(s_i, t) = \xi_k(s_i, t) + \varepsilon_{x_k}(s_i, t) \tag{2}$$

where $\eta(s_i, t)$ and $\xi_k(s_i, t)$ are two latent variables, $\varepsilon_y(s_i, t) \sim N(0, \sigma_y^2(s_i))$ and $\varepsilon_{x_k}(s_i, t) \sim N(0, \sigma_{x_k}^2(s_i))$ are the measurement and model errors, supposed to be independent. Moreover, we assume that the variances $\sigma_y^2(s_i)$ and $\sigma_{x_k}^2(s_i)$ do not depend on time and are known at each site s_i.

The relation between the two latent variables is defined by the following equation:

$$\eta(s_i, t) = \beta_0 + \boldsymbol{\gamma}_p \boldsymbol{z}(s_i) + \boldsymbol{\beta}_K \boldsymbol{\xi}(s_i, t) + \omega(s_i, t) + \varepsilon_q(s_i, t), \tag{3}$$

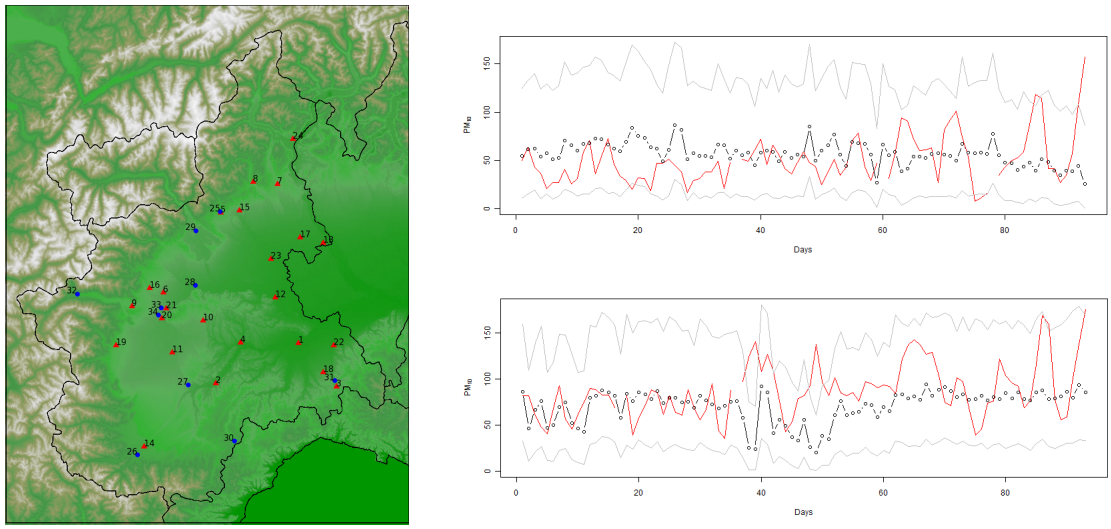
where $\boldsymbol{z}(s_i) = (z_1(s_i), \ldots, z_p(s_i))'$ is the vector of the p constant-in-time covariates known without error and $\boldsymbol{\gamma}_p = (\gamma_1, \ldots, \gamma_p)$ is the vector of their coefficients. Moreover, $\boldsymbol{\xi}(s_i, t) = (\xi_1(s_i, t), \ldots, \xi_K(s_i, t))'$ denotes the vector of the K "true" covariate values and $\boldsymbol{\beta}_K = (\beta_1, \ldots, \beta_K)$ is the vector of their coefficients. The term $\omega(s_i, t)$ is a spatio-temporal process assumed to be i.i.d. over time, so that the spatio-temporal covariance function is given by

$$Cov(\omega(s_i, t), \omega(s_j, t')) = \begin{cases} 0 & \text{if } t \neq t' \\ \sigma_\omega^2 \rho_\phi(h) & \text{if } t = t' \end{cases}$$

where $h = ||s_i - s_j||$ is the Euclidean distance between site s_i and s_j and σ^2_ω is the constant-in-time-and-space variance of the process. The function $\rho_\phi(h) = \exp(-\frac{h}{\phi})$ depends on the parameter ϕ, representing the decay rate of a spatial correlation with spatial distance. Finally, $\varepsilon_q(s_i, t)$ in Eq.(3) is the equation error that takes into account the not optimal relation between $\eta(s_i, t)$ and $\boldsymbol{\xi}(s_i, t)$; it is supposed to be normally distributed with zero mean and common variance σ^2_q. Thus, the parameter vector to be estimated is $\Phi = \{\beta_0, \boldsymbol{\gamma}_p, \boldsymbol{\beta}_K, \phi, \sigma^2_\omega, \sigma^2_q\}$. As regards inference, i.e. parameter estimation and spatial prediction of PM_{10} concentration at a new location s_0 and time t, we adopt a fully Bayesian framework via Markov chains Monte Carlo (MCMC) methods implemented through the `WinBUGS` software.

3 Results and concluding remarks

An exploratory analysis of the case study data showed skewed distributions for the considered variables. In order to make the PM_{10} and covariate distributions approximately Normal, a Box-Cox transformation (Box and Cox, 1964) was applied to the original data. Moreover, we standardized - site by site - the covariate data, in order to remove the effects related to the different ranges.



(a) PM_{10} site location (b) Time series predictions

Figure 1: Locations of the 24 PM_{10} monitoring sites (red triangles) and 10 validation stations (blue dots) and prediction of PM_{10} for *26 Borgo San Dalmazzo* (top) and *28 Chivasso* (bottom) station: solid red line refers to PM_{10} observations, black dots to PM_{10} predictions and grey solid lines to 95% prediction intervals.

With regards to the variances supposed known in the model, in this preliminary study we fixed $\sigma^2_{x_k}(s_i) = 1$, $\forall i, k$ and $\sigma^2_y(s_i) = \sigma^2_y$ where σ^2_y is the variance of

PM_{10} data all over the sites. Considering the posterior estimates for the covariate coefficients, as expected there is a significative negative relationship between PM_{10} and altitude, as well as mean wind speed, mean temperature and maximum mixing height. The posterior mean of ϕ is 90.0423 which means that the spatial correlation decreases slowly with distance: for example, at 50 km the correlation is 0.5739 and 0.1212 at 190 km. Figure 1(b) displays the predicted PM_{10} for two different validation stations (*26 Borgo San Dalmazzo* and *28 Chivasso*). It seems that the predictions are close to the observed average for each of the ten sites, even though some problems can be detected when very high or very low PM_{10} concentration levels occur in contiguous days giving rise to a higher local variability. A possible solution to this issue can be achieved by choosing different values, one per site, of PM_{10} and covariate variances, in order to take into account the possibly different measurement error of PM_{10} and numerical model error of covariates in the sites.

Moreover, our ongoing research is focused on facing the so-called "change of support problem", which arises when the numerical model output is provided at a different spatial resolution from the scale of the PM measurements. Thus, it is interesting to extend the proposed spatio-temporal model in order to deal with both point-referenced and areal data.

References

Bande S., Clemente M., De Maria R., Muraro M., Picollo M., Arduino G., Calori G., Finardi S., Radice P., Silibello C., Brusasca G. (2007) The modelling system supporting Piemonte region yearly air quality assessment. *Proceedings of 6th International Conference on Urban Air Quality, Limassol*, Cyprus, 27-29 March 2007.

Bayarri M.J., Berger J., Steinberg D.M. (2009) Special Issue on Computer Modeling, *Technometrics*, 51(4), 353-353.

Box G.E.P., Cox D.R. (1964) An analysis of transformations, *Journal of the Royal Statistical Society, B*, 26, 211-246.

Van de Kassteele J., Stein A. (2006a) A model for external drift kriging with uncertain covariates applied to air quality measurements and dispersion model output, *Environmetrics*, 17, 309-322.

Van de Kassteele J., Koelemeijer R.B.A., Dekkers A.L.M., Schaap M., Homan C.D., Stein A. (2006b) Statistical mapping of PM_{10} concentrations over Western Europe using secondary information from dispersion modeling and MODIS satellite observations, *Stochastic Environmental Research and Risk Assessment*, 21, 2, 183-194.