A Hierarchical Bayesian Spatio-Temporal Model to Estimate the Short-term Effects of Air Pollution on Human Health

Un modello bayesiano spazio-temporale gerarchico per stimare gli effetti di breve periodo dell'inquinamento atmosferico sulla salute umana

Fontanella Lara, Ippoliti Luigi and Valentini Pasquale

Abstract We introduce a hierarchical spatio-temporal regression model to study the spatial and temporal association existing between health data and air pollution. The model is developed for handling measurements belonging to the exponential family of distributions and allows the spatial and temporal components to be modelled conditionally independently via random variables for the (canonical) transformation of the measurements mean function. A temporal autoregressive convolution with spatially correlated and temporally white innovations is used to model the pollution data. This modelling strategy allows to predict pollution exposure for each district and afterwards these predictions are linked with the health outcomes through a spatial dynamic regression model.

Abstract *In questo lavoro viene introdotto un modello di regressione spazio-temporale gerarchico per studiare l'associazione spazio-temporale tra i dati sulla salute e l'inquinamento atmosferico. Il modello è stato sviluppato nell'ambito della famiglia esponenziale e consente di modellare condizionatamente le componenti spaziali e temporali in modo indipendente attraverso variabili casuali per la trasformazione (canonica) della funzione media. Per modellare i dati sull'inquinamento viene utilizzata una convoluzione temporale autoregressiva con innovazioni spazialmente correlate e temporalmente indipendenti. Questa strategia di modellizzazione consente di prevedere l'esposizione all'inquinamento per ciascun distretto e, successivamente, queste previsioni sono messe in relazione con le ospedalizzazioni attraverso un modello di regressione spaziale dinamica.*

Key words: Hierarchical model, spatio-temporal model, MCMC

Fontanella Lara
DSGS, University "G.d'Annunzio" of Chieti-Pescara, e-mail: lara.fontanella@unich.it

Ippoliti Luigi
DeC, University "G.d'Annunzio" of Chieti-Pescara e-mail: luigi.ippoliti@unich.it

Valentini Pasquale
DeC, University "G.d'Annunzio" of Chieti-Pescara e-mail: pasquale.valentini@unich.it

1 Introduction

In the last 30 years, a large number of studies have provided substantial statistical evidence of the adverse health effects associated with air pollution. The statistical literature on health care research is very rich and includes a plethora of models referring to different types of study designs. Most of those studies are usually based on time series models, developed both in single and multisites frameworks (e.g., [1]). However, because air pollution concentrations vary at fine spatio-temporal scales, quantifying the impact of air pollution appeares more as an inherently spatio-temporal problem. Also, despite the availability of large data sets for multiple pollutants, only a few studies consider the joint effects of numerous air pollutants simultaneously [2]. In this paper, we thus propose a hierarchical spatio-temporal regression model, which is able to cope with different spatial resolutions in order to change the support of air pollution data (regressors) to achieve alignment with the health outcome measured at area level.
Considering the air pollution data, we opt for a modelling approach based on assuming the existence of a latent Gaussian variable which may be interpreted as a potential pollution harmful to health in the short period. Then at the process level, we use a dynamic model proposed by [3] which takes into account for spatio- temporal variation using a temporal autoregressive variable with spatially correlated innovations. It is assumed that these innovations follow a Gaussian process with an exponential covariance function. Given such model, we can interpolate the process at unobserved times and/or locations and face the change of support problem (COSP). Afterwards, assessing the effect of air pollution on human health is possible through a regression model that includes lagged exposure variables as covariates.

2 Model Specification

Assume that Y and X are two multivariate spatio-temporal processes observed at temporal instants $t = 1,2,\ldots,T$ and generic locations, $\mathbf{s} \in \mathscr{D}_y$ and $\mathbf{u} \in \mathscr{D}_x$, respectively. Assume also that X is a predictor of Y, which thus represents the process of interest. For the two different processes, the spatial sites $\mathbf{s}$ and $\mathbf{u}$ can denote the same location but, in general, they need not be the same. Furthermore, both $\mathscr{D}_y$ and $\mathscr{D}_x$ may represent different spatial characteristics and structures. Usually, health data (Y) are collected over time in a fixed study region,$\mathscr{D}_y$, typically in the form of mortality and morbidity counts or hospital admissions, coded according to the type of disease (e.g. cardiovascular, acute respiratory, etc). While pollution concentrations (X) are measured at specific points in time and at a number of monitoring sites across a continuous region $\mathscr{D}_x$ and usually come in the form of geostatistical data.

Let n_y be the number of observed variables for Y and n_x the number of observed variables for X. The most informative case is represented by the isotopic configuration where, for each multivariate process, Y or X, all variables are measured at all their respective sites. In this case, let $\mathbf{Y}(\mathbf{s},t) = [Y_1(\mathbf{s},t),\ldots,Y_{n_y}(\mathbf{s},t)]'$

be the vector of the n_y values of Y at site $\mathbf{s}$ and time t. Equivalently, we write $\mathbf{X}(\mathbf{u},t) = [X_1(\mathbf{u},t),\ldots,X_{n_x}(\mathbf{u},t)]'$ for the vector of the n_x values of X at site $\mathbf{u}$ and time t. The opposite case is the completely heterotopic case where not all the variables can be observed at the same site − this is especially true for X in our study. Without loss of generality, for the sake of simplicity, here we use the notation for the isotopic case. Accordingly, the n_y variables of Y are observed at the same sites $\mathbf{s}_i,\ i = 1,\ldots,N_y$ and the n_x variables of X are observed at sites $\mathbf{u}_r,\ r = 1,\ldots,N_x$.

Let $\tilde{n}_y = n_y N_y$ and $\tilde{n}_x = n_x N_x$. At a specific time t, by using a site ordering, the $(\tilde{n}_y \times 1)$ and $(\tilde{n}_x \times 1)$ dimensional spatial processes are denoted as $\mathbf{Y}(t) = [\mathbf{Y}(\mathbf{s}_1,t)',\ldots,\mathbf{Y}(\mathbf{s}_{N_y},t)']'$ and $\mathbf{X}(t) = [\mathbf{X}(\mathbf{u}_1,t)',\ldots,\mathbf{X}(\mathbf{u}_{N_x},t)']'$. However, the data may also be ordered by variable. In this case, we write $\mathbf{Y}(t) = [\mathbf{Y}_1(t)',\ldots,\mathbf{Y}_{n_y}(t)']'$ and $\mathbf{X}(t) = [\mathbf{X}_1(t)',\ldots,\mathbf{X}_{n_x}(t)']'$, where $\mathbf{Y}_k(t)$ is the the vector of n_y observations for variable Y_k, and $\mathbf{X}_j(t)$ is the the vector of n_x observations for variable X_j.

The model is based on the measurement equations for the conditionally independent variables,

$$Y_k(\mathbf{s},t)|\eta_{y_k}(\mathbf{s},t),\sigma^2_{y_k} \overset{ind}{\sim} F_y(\eta_{y_k}(\mathbf{s},t),\sigma^2_{y_k}), \quad k = 1,\ldots,n_y$$

$$X_j(\mathbf{u},t)|\eta_{x_j}(\mathbf{u},t),\sigma^2_{x_j} \overset{ind}{\sim} F_x(\eta_{x_j}(\mathbf{u},t),\sigma^2_{x_j}), \quad j = 1,\ldots,n_x,$$

where $\sigma^2_{y_k}$ and $\sigma^2_{x_k}$ are dispersion parameters. In general, the distributions F_y and F_x are allowed to be from any exponential family distribution. By choosing appropriate canonical link functions, the specification of the measurement equations are completed with the specification of the following linear predictors

$$g_y\left[\eta_{y_k}(\mathbf{s},t)\right] = \mu_{y_k}(\mathbf{s},t) + \phi_{y_k}(\mathbf{s},t) \tag{1}$$

$$g_x\left[\eta_{x_j}(\mathbf{u},t)\right] = \mu_{x_j}(\mathbf{u},t) + \phi_{x_j}(\mathbf{u},t) \tag{2}$$

where $\mu_{y_k}(\mathbf{s},t)$ and $\mu_{x_j}(\mathbf{u},t)$ are fixed effect terms representing the large-scale spatio-temporal variability of the processes, and $\phi_{y_k}(\mathbf{s},t)$ and $\phi_{x_j}(\mathbf{u},t)$, are random effects introduced to capture any residual spatio-temporal autocorrelation.

The random effects are modelled through the following equations

$$\phi_{x_j}(\mathbf{u},t) = \int_{\mathscr{D}_x} \kappa_{\theta_{x_j}}(\mathbf{u}-\mathbf{u}')\phi_{x_j}(\mathbf{u}',t-1)d\mathbf{u}' + \mathbf{v}_{x_j}(\mathbf{u},t) \tag{3}$$

$$\phi_{y_k}(\mathbf{s},t) = \sum_{j=1}^{n_x}\sum_{l=0}^{L} \beta_{k,j,l}\ \phi_{x_j}(\mathbf{s},t-l) + \mathbf{v}_{y_k}(\mathbf{s},t) \tag{4}$$

where $\kappa_{\theta_{x_j}}(\mathbf{u}-\mathbf{u}') = \rho_{1,x_j}\exp\left((\mathbf{u}-\mathbf{u}')'\,\Sigma^{-1}_{x_j}\,(\mathbf{u}-\mathbf{u}')\right)$,
$\Sigma^{-1}_{x_j} = \frac{1}{\rho^2_{2,x_j}}\begin{bmatrix}\cos(\alpha_{x_j}) & \sin(\alpha_{x_j})\\ -d_{x_j}\sin(\alpha_{x_j}) & d_{x_j}\cos(\alpha_{x_j})\end{bmatrix}$, $\alpha_{x_j} \in [0,\frac{\pi}{2}]$, $d_{x_j} > 0$, $\theta_{x_j} = \{\rho_{1,x_j},\rho_{2,x_j},c_{x_j},\alpha_{x_j}\}$,
$\beta_{k,j,l}$ is the distributed lag coefficient which relates the jth pollutants at lag l to the kth disease health outcome, $\mathbf{v}_{y_k}(\mathbf{s},t)$ and $\mathbf{v}_{x_j}(\mathbf{u},t)$ Gaussian innovations that are

white in time and correlated in space.
It is worth noting that $\phi_{x_j}(\mathbf{s},t-l)$, in equation (4), is the structured variation in time at space resolution $\mathbf{s}$, $\phi_{x_j}(\mathbf{s},t) = \int_{\mathbf{s}} \phi_{x_j}(\mathbf{u},t)d\mathbf{u}$. In practice one could first define a regular grid, then interpolate the non-observed grid points, and approximate the integral by a Riemann sum. Since the regular grid usually becomes very large, this is computationally expensive, other strategies to approximate the integral can be found in [3].
Model completion requires specific forms for $\mu_y(t)$ and $\mu_x(t)$. The simplest specification of the mean components assumes the form of a linear regression function to take care of the effects of confounders, i.e. $\mu_{x_l}(\mathbf{u},t) = \sum_{i=1}^{c}\sum_{l=0}^{g} \delta_{x_j,il}(\mathbf{u})\, z_i(\mathbf{u},t-l)$, and $\mu_{y_k}(\mathbf{s},t) = \sum_{i=1}^{c}\sum_{l=0}^{g} \delta_{y_k,ij}(\mathbf{s})\, z_i(\mathbf{s},t-l)$, where $z_i(\cdot,t),\ i=1,\ldots,c$ are observed covariates or components representing seasonal and long-term trends introduced to take care of the effects of unmeasured confounders (see [1] and [4]). Note that the $z_i(\cdot,t)$ could also be smoothed versions of measured confounders represented by natural cubic splines with specified degree of freedom.
The hierarchy of our model is completed by specifying the prior distributions of all hyperparameters. Noninformative conjugate priors are assumed with the expection of the spatial correlation parameters. This model is developed within a state-space framework and full probabilistic inference for the parameters is facilitated by a Markov chain Monte Carlo (MCMC) scheme for multivariate dynamic systems.
The proposed model has an intuitive appeal and enjoys several advantages. For example, it describes the spatial-temporal variability of the disease risk and explicitly defines a non-separable spatio-temporal covariance structure of the process. Also, it allows to study how the disease risk at a specific areal unit reacts over time to exogenous impulses from the same or different areal units. Finally, several general structures that make use of different covariate information, can be easily accommodated in the different levels of the hierarchy.
Fitting our model using the MCMC algorithm is computationally intensive. However, once the statistical model is fitted and assuming that the posterior of the parameters for (3) does not change (see [3] for more details), predictions are computationally a lot cheaper.

3 Application

We illustrate our modelling approach to measure the effect of exposure variables on hospital admissions observed in Lombardia and Piemonte regions (Italy) in 2011. In particular, health data consist of counts of daily hospital admissions for cardiovascular diseases and respiratory diseases. Pollution data refer to daily-average concentration levels of CO, NO_2, PM_{10} and O_3.
To provide more insight on the way in which the disease risks spread out to surrounding districts, Figure 1 below shows maps of raw standardized morbidity ratios obtained by averaging the SMR values across time. The map on the left, shows that, on average, the highest risk areas associated with the cardiovascular diseases

correspond to districts in the Southeastern parts of Lombardia. The map for the respiratory diseases (right) also supports the idea that Lombardia is the most at risk with the highest SMR values observable at the Northern and the Southeastern parts of Milan. In general, the SMR maps show evidence of localised spatial clusters.
The MCMC algorithm was run for $35,000$ iterations. Posterior inference was based on the last $30,000$. Convergence was monitored by inspecting trace plots. Preliminary results show a positive association between air pollutants and hospital admissions. In particular, the peak response of hospital admissions for cardio-respiratory diseases after a positive shock on pollutants occurs after three days and then gradually decreases and dies out in about six days.

Acknowledgements This work is developed under the PRIN2015 supported project 'Environmental processes and human activities: capturing their interactions via statistical methods (EPHASTAT)' [grant number 20154X8K23] funded by MIUR (Italian Ministry of Education, University and Scientific Research).

References

1. Peng, R.D. and Dominici, F. and Louis, T.A.: Model choice in time series studies of air pollution and mortality. J. R. Stat. Soc. Ser. A. Stat. Soc., **169**, 179–203 (2006)
2. Rushworth, A. and Lee, D. and Mitchell, R.: A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. Spat. Spatiotemporal. Epidemiol. **10**, 29–38 (2014)
3. Sigrist, F., Knsch, H.R., Stahel, W.A.: A dynamic nonstationary spatio-temporal model for short term prediction of precipitation. Ann. Appl. Stat.,**6**, 1452–1477 (2012)
4. Shaddick, G. and Zidek, J.: Spatio-Temporal Methods in Environmental Epidemiology, Chapman Hall/CRC (2015)

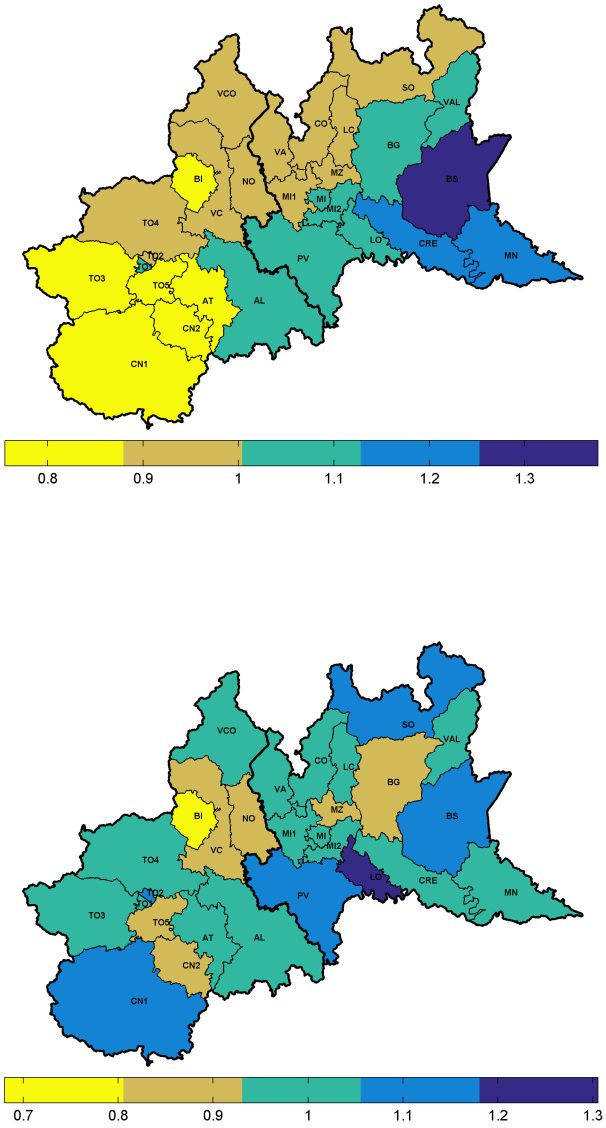

Fig. 1 *Map of the standardised morbidity ratio (SMR) for hospital admissions due to cardiovascular (left) and respiratory (right) diseases in 2011. The lable within each district represents the acronymous of the ASL.*