# A comparison between hierarchical spatio-temporal models in presence of spatial homogeneous groups: the case of Ozone in the Emilia-Romagna Region[1]

Francesca Bruno, Lucia Paci
Dipartimento di Scienze Statistiche, Università di Bologna, lucia.paci2@unibo.it

**Abstract**: Hierarchical spatio-temporal models permit to estimate many sources of variability. In many environmental problems, different features characterizing spatial locations can be found. Differences in these classifications can show discrepancies either in mean levels or in the spatio-temporal dependence structure. When these characteristics are not included in the model structure, model performances and spatial predictions may lead to poor results. Here, we compare alternative enrichments of the hierarchical spatio-temporal model that consider the presence of groups. Our application concerns Ozone data in the Emilia-Romagna region in which the monitoring sites can be classified according to their relative position with respect to traffic emissions.

**Keywords**: spatio-temporal models, hierarchical models, groups of sites, ozone data.

## 1. Introduction

Hierarchical models, being very flexible, are suitable for dealing with differences both at the measurement and the process level (Wikle, 2003).
In the following, we expand the general framework describing hierarchical spatio-temporal models for studying geostatistical data by the inclusion of domain classifications with respect to certain differentiating features (Wang *et al*., 2009). When studying air pollution, for example, monitoring stations may be differently located with respect to traffic or household density. This peculiarity can be modeled in a number of different ways. Models that allow for differences between groups of sites have recently been proposed (Cocchi and Bruno, 2010). In environmental applications: for example, Paci (2010) proposed a hierarchical spatio-temporal model for pollutants where the group differences were captured by the intercept of the model (*i.e.* difference in pollution levels between the urban and rural locations). In Sahu *et al*. (2006) a hierarchical space-time model for $PM_{2.5}$ that includes two spatio-temporal processes was proposed, where the first captures the background effects, and the second adds extra variability for urban locations by using the relationship between the response variable and suitable covariates (the population density, in this case).
Here the inclusion of groups in spatio-temporal models is formalized in a more general way. We describe alternative proposals for including group differences in hierarchical Bayesian models. The assessment of the consequences for spatial prediction under this innovation will be also considered.

[1] Work supported by the project PRIN 2008: New developments in sampling theory and practice, Project number 2008CEFF37, Sector: Economics and Statistics, awarded by the Italian Government.

This paper is organized as follows: the next section describes the Ozone dataset; Section 3 sketches the main models that include spatial groups; the final section presents the main results and some concluding remarks.

## 2. The Ozone Dataset

Tropospheric ozone is one of the most important pollutants when studying air quality. Here, the dataset consists of Ozone daily measurements (in $\mu g/m^3$) collected from 31 monitoring stations across the Emilia–Romagna Region in 2001. Monitoring sites can be classified according to traffic pollution exposure (D.M.A. 16/05/1996); the two groups consist of 17 background monitoring sites (denoted by "G1") and 14 sites characterized by their vicinity to traffic emissions (denoted by "G2"). Monitoring sites belonging to G1 are expected to measure higher Ozone levels than sites belonging to G2. Some meteorological covariates are available for each site and each time. In particular, one of the most correlated with Ozone is the daily mixing height, that will be included as a covariate in the model.

## 3. Model specification

Let $\mathbf{Y}^* = \{\mathbf{Y}^*(\mathbf{u}, t); \mathbf{u} \in (\mathbf{u}_1, \cdots, \mathbf{u}_{n^*}), t \in (1, \cdots, T)\}$ denote the log-Ozone concentrations for the generic location and time $(\mathbf{u}, t)$. We consider 27 of the 31 sites for estimation and 4 sites for prediction assessment (2 for each group). Let define $\mathbf{Y}$ as the $Tn$-dimensional subset of the original dataset under these specifications.
Following the usual hierarchical spatio-temporal specification (Banerjee *et al*. 2004), let

$$\mathbf{Y} = \mathbf{Z} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{Z}$ is the $Tn$-dimensional spatio-temporal process and $\boldsymbol{\varepsilon}$ is a Gaussian noise process $N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_{Tn \times Tn})$, representing the spatio-temporal measurement error structure via homoscedastic and independent components. Conditionally on $\mathbf{Z}$ and $\sigma_\varepsilon^2$ the distribution of $\mathbf{Y}$ is:

$$\mathbf{Y}|\mathbf{Z}\,,\ \sigma_\varepsilon^2 \sim N(\mathbf{Z}\,, \sigma_\varepsilon^2 \mathbf{I}_{Tn \times Tn})$$

The second stage of the hierarchy can be defined as the combination of a large scale spatio-temporal process ($\mathbf{m}$), a spatial effect ($\mathbf{W}$) and a temporal effect ($\mathbf{V}$):

$$\mathbf{Z} = \mathbf{m} + \mathbf{1}_{T \times 1} \otimes \mathbf{W} + \mathbf{V} \otimes \mathbf{1}_{n \times 1} \tag{2}$$

The expression for the $Tn$-dimensional trend component ($\mathbf{m}$) is:

$$\mathbf{m} = \mathbf{X}\boldsymbol{\beta} \tag{3}$$

where $\boldsymbol{\beta} = (\beta_0\,, \beta_1)'$ and $\mathbf{X}$ is a $Tn \times 2$ covariates matrix with unit values in the first column and daily mixing heights in the second column. The expression in (2) provides additive temporal and spatial effects (multiplicative on the original scale). The temporal random effect $\mathbf{V} = \big(V(1)\,, \cdots, V(T)\big)'$ and the spatial random effect $\mathbf{W} = \big(W(\mathbf{u}_1), \cdots, W(\mathbf{u}_n)\big)'$ capture respectively any spatial and temporal dependence which remains unexplained by the model for the mean (3). The distribution of the random effect $\mathbf{V}$ can be expressed via the multivariate distribution

$$\mathbf{V} \sim N\big(\mathbf{0},\ \sigma_v^2\, \mathbf{A}(\phi)\big) \quad \text{where} \quad \big(\mathbf{A}(\phi)\big)_{ij} = exp\big(-\phi\, \|t_i - t_j\|\big) \tag{4}$$

and $\sigma_v^2$ is the scalar variance of the temporal component; $\mathbf{A}(\phi)$ is the $T \times T$ correlation matrix defined by the exponential function.
The spatial random effect $\mathbf{W}$ is modeled as a Gaussian process

$$\mathbf{W} \sim N\left(\mathbf{0},\ \sigma_w^2\ \mathbf{H}(\delta)\right) \quad \text{where} \quad \left(\mathbf{H}(\delta)\right)_{ij} = exp\left(-\delta \left\|\mathbf{u}_i - \mathbf{u}_j\right\|\right) \tag{5}$$

and $\sigma_w^2$ is the scalar variance of the spatial process; $\mathbf{H}(\delta)$ is the $n \times n$ spatial exponential correlation matrix.
The model hierarchy is completed by the specification of noninformative prior distributions for the hyperparameters.
In the following subsections we propose two different specifications of model (1) – (5) (from now on called "Model (A)") in order to take groups into account.

*3.1 Modeling differences in the trend component*

When the differences between the two groups are captured by the average level, the discrepancies are developed from model (3), the large-scale process can be rewritten as:

$$\mathbf{m} = \alpha \mathbf{d}_{\mathbf{u} \in G1,(t=1,\cdots,T)} + \mathbf{X}\boldsymbol{\beta} \tag{6}$$

In (6) $\alpha$ is a scalar type-specific intercept and $\mathbf{d}_{\mathbf{u} \in G1,(t=1,\cdots,T)}$ is a *Tn*-dimensional vector collecting the dummy variables that classify the spatial sites into groups. The $\beta_0$ parameter represents the intercept for the sites belonging to G2 and $\alpha + \beta_0$ represents the intercept for the other group. This model will be referred to as "Model (B)".

*3.2 Modeling differences in the spatio-temporal covariance structure*

When differences in the spatio-temporal dependence structure are included in the model, alternative $\sigma_w^2\ \mathbf{H}(\delta)$ might be considered in (5). Matrix $\sigma_w^2\ \mathbf{H}(\delta)$ is constituted by blocks, with group-specific spatial variance matrices in the diagonal after reordering sites according to the groups. The most complex model includes an out-of-diagonal between-group variance block matrix, $\sigma_{w(G1,G2)}^2 \mathbf{K}\left(\delta_{G1,G2}\right)$, that is characterized by group parameters:

$$\sigma_w^2\ \mathbf{H}(\delta) = \begin{bmatrix} \sigma_{w(G1)}^2\ \mathbf{H}(\delta_{G1}) & \sigma_{w(G1,G2)}^2\ \mathbf{K}\left(\delta_{G1,G2}\right) \\ \sigma_{w(G1,G2)}^2\ \mathbf{K}\left(\delta_{G1,G2}\right) & \sigma_{w(G2)}^2\ \mathbf{H}(\delta_{G2}) \end{bmatrix} \tag{7}$$

Specification (7) needs the estimation of a huge number of parameters. When interactions between locations belonging to different groups are ignored, $\sigma_{w(G1,G2)}^2 \mathbf{K}\left(\delta_{G1,G2}\right)$ is fixed at zero (in what follows "Model (C)").

## 3. Results

The comparison between models is performed both in terms of goodness of fit (via DIC) and in terms of predictive assessment (via Predictive Model Choice Criterion, PMCC, Sahu *et. al* 2006).Table 1 shows that Model (B) has the best performance. This highlights that the main differences between groups concern the mean levels and it is reasonable to assume a common correlation structure for both groups.

| | Model (A) for G1 sites | Model (A) for G2 sites | Model (A) for all sites | Model (B) | Model (C) |
|---|---|---|---|---|---|
| DIC | 6403 | 5737 | 11840 | 11830 | 11840 |
| PMCC | 187.50 | 231.03 | 391.32 | 377.13 | 400.99 |

**Table 1**: DIC and PMCC for all models considered

Figure 1 shows the predictions for a specific site and for all models. The predictive performances are similar for all models, the prediction credibility bands contain almost always the observed values.

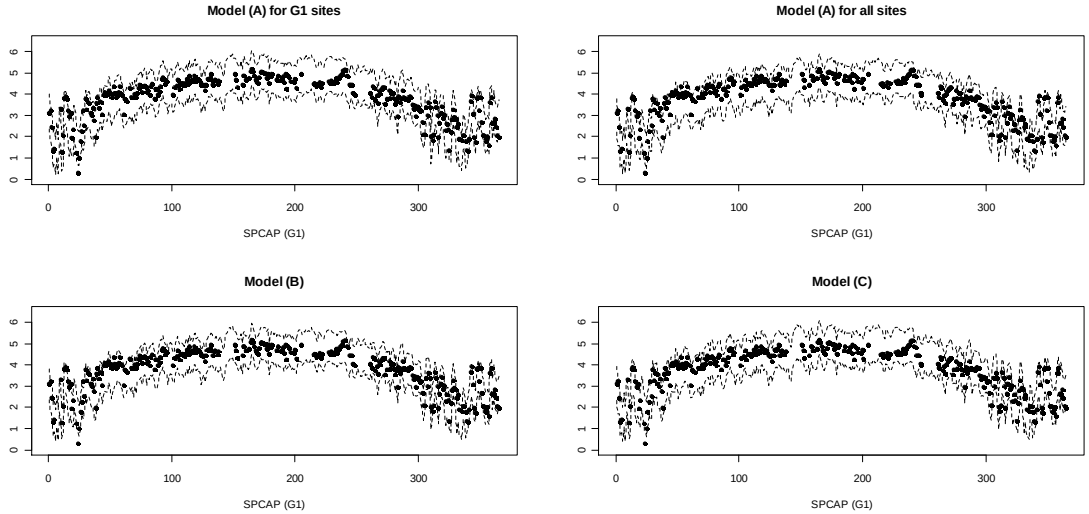


**Figure 1**: Predictions for a site belonging to G1 for 2001, estimated for all models

## References


Banerjee S., Carlin B.P., Gelfand A.E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall, CRC Press.

Cocchi D., Bruno F. (2010) Considering groups in the statistical modeling of spatio-temporal data, *Statistica*, 4, *in press*.

D.M.A. (16/05/1996) Attivazione di un sistema di sorveglianza di inquinamento da Ozono. In Italian.

Paci L. (2010) Hierarchical Bayesian space-time model: the case of Ozone in Emilia Romagna, Thesis of master in statistics (in Italian).

Sahu S.K., Gelfand A.E., Holland D.M. (2006) Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 61-86.

Wang J., Christakos G., Hu M-G. (2009) Modeling spatial means of surfaces with stratified nonhomgeneity, *IEEE Transactions on Geoscience and Remote Sensing*, 47, 4167-4174.

Wikle C.K. (2003) Hierarchical models in environmental science, *International Statistical Review* 71, 181-199.