Spatio-temporal modelling in the presence of few time points

L. Margalho[1,*], R. Menezes[2] I. Sousa[1]

[1] *Department of Physics and Mathematics-Coimbra Institute of Engineering and CMAT; lmelo@isec.pt*
[2] *Department of Mathematics and Applications-University of Minho and CMAT; rmenezes@math.uminho.pt*
[3] *Department of Mathematics and Applications-University of Minho and CMAT; isousa@math.uminho.pt*
[*] *Corresponding author*

Abstract. *Environmental monitoring networks are providing large amounts of spatio-temporal data. Air pollution data, as other environmental data, exhibit a spatial and a temporal correlated nature. To improve the accuracy of predictions at unmonitored locations, there is a growing need for models capturing those spatio-temporal correlations.*
With this work, we propose a spatio-temporal model for gaussian data collected in a few number of surveys. We assume the spatial correlation structure to be the same in all surveys. Moreover, as a consequence of the reduced number of time observations, the temporal correlations are modeled as fixed effects. A simulation study, aiming to validate the model, is conducted. The proposed model is applied to heavy metal concentration data, collected using moss biomonitors in Portugal, from three surveys which occurred between 1992 and 2002. Prediction maps of the observed variable for the most recent survey, together with the corresponding prediction variance as an uncertainty measure, are presented.

Keywords. *Air pollution; Spatio-temporal data; Spatial and temporal correlations; Separability.*

1 Introduction

The Portuguese participation in the international mapping project *Atmospheric Heavy Metal Deposition in Europe* yielded concentration values of several heavy metals in biomonitoring moss samples. In the literature, one may found several spatio-temporal models applied to monitoring data. Cocchi *et al.* (2007), under the Bayesian framework, use data from 11 spatial locations collected over 1096 days. Bruno *et al.* (2003) use a data set consisting of daily ozone measurements made at 32 monitoring locations, for the period 1998-2002, enabling the identification of the temporal variability which, when removed, leaves separable space and time correlation components. Mitchell *et al.* (2005), aiming to study the effect of high levels of CO_2 on rice using data from 13 spatial locations and 112 time points, test the separability

of the proposed spatio-temporal model by rearranging the data as in the context of multivariate repeated measures. These examples, opposite to the one here proposed, share one common feature: the number of time observations is (much) larger than the number of spatial locations.
There are examples, however, where data are collected over a large number of spatial locations but only few times, disabling the use of time series techniques. Margalho *et al.* (2014) proposed an extension of an existing spatio-temporal model, using the previously mentioned portuguese biomonitoring data.

2 The model

We propose a spatio-temporal model for Gaussian data, collected at location $\mathbf{s}$ and time t,

$$Y(\mathbf{s},t) = \mu(\mathbf{s},t) + Z(\mathbf{s},t) + \varepsilon(\mathbf{s},t) \tag{1}$$

The mean component $\mu(\mathbf{s},t)$, depending on possibly observed covariates $f_i(\mathbf{s},t)$, will be considered as

$$\mu(\mathbf{s},t) = \sum_{i=1}^{p} \beta_i f_i(\mathbf{s},t) \tag{2}$$

where $E[Y(\mathbf{s},t)] = \mu(\mathbf{s},t)$. The non-observed spatio-temporal process $Z(\mathbf{s},t)$ is such that

$$Z(\mathbf{s},t) \sim MVN(0,\Sigma) \tag{3}$$

and $\varepsilon(\mathbf{s},t)$ represents gaussian space-time measurements errors,

$$\varepsilon(\mathbf{s},t) \sim N(0,\tau^2) \tag{4}$$

In the space-time process (3), considering N locations observed at T surveys, Σ is represented by a $T \times T$ symmetric matrix whose elements Σ^{t_k,t_l} are $N \times N$ matrices, where the element on line i and column j is

$$\Sigma_{ij}^{t_k,t_l} = \mathrm{Cov}\left[Z(\mathbf{s}_i,t_k), Z(\mathbf{s}_j,t_l)\right] \ , \ k,l = 1,\cdots,T; \ i,j = 1,\cdots,N \tag{5}$$

The proposed model assumes an isotropic and separable covariance structure, so we define purely spatial and purely temporal covariance functions, Cov_S and Cov_T, resulting in

$$\begin{aligned} \mathrm{Cov}\left[Z(\mathbf{s}_i,t_k), Z(\mathbf{s}_j,t_l)\right] &= \mathrm{Cov}_S\left(\|\mathbf{s}_i - \mathbf{s}_j\|\right) \times \mathrm{Cov}_T\left(|t_k - t_l|\right) \\ &= \mathrm{Cov}_S\left(h_S\right) \times \mathrm{Cov}_T\left(h_T\right) \end{aligned} \tag{6}$$

Under the assumption of second order stationarity, we propose two different interpretations for the covariance function. Denoting by σ_S^2 the spatial variance and by σ_T^2 the temporal variance,

$$\Sigma(h_S,h_T) = \sigma_S^2 R_S(h_S) \circ \sigma_T^2 R_T(h_T) \tag{7}$$

As an alternative, and denoting by σ_{total}^2 the overall variance,

$$\Sigma(h_S,h_T) = \sigma_{total}^2 R_S(h_S) \circ R_T(h_T) \tag{8}$$

($\circ$ represents elementwise product of matrices).

3 Simulation study

For a model validation purpose, a simulation study was conducted. A set of 50 randomly chosen space locations was considered in the square $[0,1]^2$. In order to have a region with more intensified sampling density, mimicking the behavior of the real data set used in the application, 15 of those locations belong to the square $[0.45,0.55]^2$. Observations are assumed to be collected at 3 different moments, according to an AR(2) model. The mean component (2) includes the covariates *intensity of the sampling design*, $int(\mathbf{s})$, and the specific contribution of a given survey, $v_i(t)$, resulting in

$$\mu(\mathbf{s},t) = \beta_0 + \beta_1 int(\mathbf{s}) + \beta_2 v_2(t) + \beta_3 v_3(t)$$

where

$$v_i(t) = \begin{cases} 1 & \text{if } t = i \\ & \qquad\qquad i = 2,3 \\ 0 & \text{otherwise} \end{cases}$$

Using the absolute value of the coefficient of variation as an accuracy measure, simulations with $\Sigma(h_S, h_T)$ given by (7) provided better results than when using (8).

4 Application

The model described in (1) assumes that the hidden process $Z(\mathbf{s},t)$ and the measurement error $\varepsilon(\mathbf{s},t)$ are Gaussian ((3) and (4)). It is well known (*e.g.* Cressie and Wikle (2011)), that for a non-observed location $\mathbf{s}_0$ and a time t_0, the joint distribution of $Y(\mathbf{s}_0,t_0)$ and $Y(\mathbf{s},t)$ is

$$\begin{bmatrix} Y(\mathbf{s}_0,t_0) \\ \mathbf{Y}(\mathbf{s},t) \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \mu(\mathbf{s}_0,t_0) \\ \mu(\mathbf{s},t) \end{bmatrix}, \begin{bmatrix} C_{0,0} & \mathbf{c}_0^T \\ \mathbf{c}_0 & \mathbf{C}_Y \end{bmatrix} \right) \tag{9}$$

where $\mu(\mathbf{s}_0,t_0) = \widehat{\beta}_0 + \widehat{\beta}_1 int(\mathbf{s}_0) + \widehat{\beta}_3$ and $\mu(\mathbf{s},t)$ are defined by (2), $C_{0,0} = \text{Var}(Y(\mathbf{s}_0,t_0))$, $\mathbf{c}_0 = \text{Cov}(Y(\mathbf{s}_0,t_0), Y(\mathbf{s},t))$, and $\mathbf{C}_Y = \Sigma + \tau^2 I$ with Σ as in (5).
Under the assumption (9), the predicted value at an unsampled location $Y^*(\mathbf{s}_0,t_0)$ is given (Cressie and Wikle (2011)) by

$$Y^*(\mathbf{s}_0,t_0) = \text{E}[Y(\mathbf{s}_0,t_0)|\mathbf{Y}] = \mu(\mathbf{s}_0,t_0) + \mathbf{c}_0^T \mathbf{C}_Y^{-1} (\mathbf{Y}(\mathbf{s},t) - \mu(\mathbf{s},t)) \tag{10}$$

and the variance of the prediction is

$$\sigma^2(\mathbf{s}_0,t_0) = \text{E}[Y(\mathbf{s}_0,t_0) - Y^*(\mathbf{s}_0,t_0)]^2 = C_{0,0} - \mathbf{c}_0^T \mathbf{C}_Y^{-1} \mathbf{c}_0 \tag{11}$$

The estimates of the model parameters are in Table 1. The computation of the standard errors was made via Monte-Carlo simulation.

Param.	β_0	β_1	β_2	β_3	ρ_{12}	ρ_{13}	ρ_{23}	σ_S^2	σ_T^2	τ^2	ϕ
Estim.	7.449	0.007	0.152	-0.241	0.973	0.909	0.965	0.979	1.483	1.023	58624.08
St. Error	0.042	0.003	0.014	0.019	0.008	0.009	0.009	0.011	0.018	0.013	1470.20

Table 1: Model parameter estimates with standard errors, for Scenario 1

Figure 1 shows the (Box-Cox transformed) predicted concentration map for the most recent survey and the corresponding interpolation error map, over mainland Portugal.

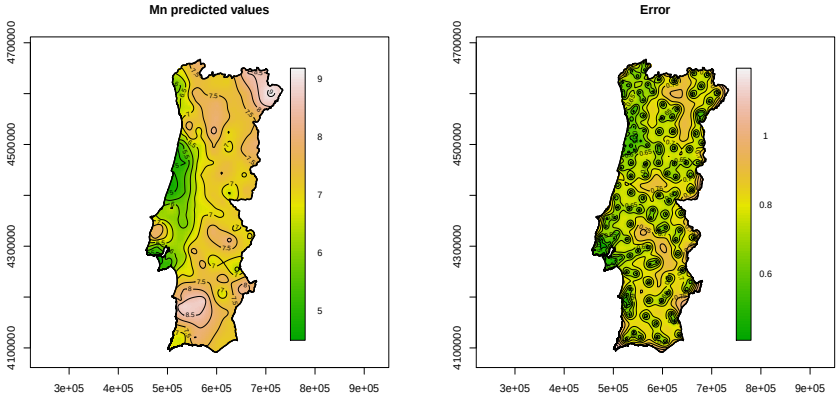

Figure 1: Mn prediction map (left) and interpolation error map (right).

Acknowledgments. The authors acknowledge financial support from the project PTDC/MAT/112338/2009 (FEDER support included) of the Portuguese Ministry of Science, Technology and Higher Education. This research was also financed by FEDER Funds through "Programa Operacional Factores de Competitividade COMPETE" and by Portuguese Funds through FCT-"Fundação para a Ciência e a Tecnologia", within the Project PEst-OE/MAT/UI0013/2014

References

[1] Bruno F., Guttorp P., Sampson P., Cocchi D. (2003) *Non-separability of space-time covariance models in environmental studies.* In Jorge Mateu, David Holland, Wenceslao González Manteiga (eds), The ISI International Conference on Environmental Statistics and Health, Universidade de Santiago de Compostela, 153–161.

[2] Cocchi D., Greco F., Trivisano C. (2007) *Hierarchical space-time modelling of PM_{10} pollution.* Atmospheric Environment, **41**, 532–542.

[3] Cressie N, Wikle K (2011) *Statistics for spatio-temporal data.* Revised edition. John Wiley & Sons Inc, Hoboken, New Jersey.

[4] de Cesare L., Meyers D.E., Posa D. (2001) *Estimating and modeling space-time correlation structures*, Statistics and Probability Letters, **51**, 9–14.

[5] Margalho L., Menezes R., Sousa I. (2014) *Assesssing interpolation errors* Stochastic Environmental Research and Risk Assessment, **28**, 1307–1321

[6] Mitchell M.,Genton M., Gumpertz M. (2005) *Testing for separability of space-time covariances.* Environmetrics, **16**: 819–831

[7] Sherman, M. (2011) *Spatial statistics and spatio-temporal data. Covariance functions and directional properties.* John Wiley & Sons, Ltd