The dynamic coregionalization model in air quality risk assessment

Francesco Finazzi
University of Bergamo, Dept. of Information Technology and Mathematical Methods
viale Marconi, 5
Dalmine 24044, Italy
E-mail: francesco.finazzi@unibg.it

Alessandro Fassò
University of Bergamo, Dept. of Information Technology and Mathematical Methods
viale Marconi, 5
Dalmine 24044, Italy
E-mail: alessandro.fasso@unibg.it

Marian E. Scott
University of Glasgow, School of Mathematics and Statistics
15 University Gardens
Glasgow G12 8QW, Scotland
E-mail: Marian.Scott@glasgow.ac.uk

1. Introduction

Air pollution monitoring and mapping at country level are challenging tasks due to the large spatial scale and the amount of data involved. Nevertheless, they can be carried out successfully by considering the proper space-time statistical models and mapping techniques, which provide estimates of both the pollutant concentration and the respective uncertainty (Fassò and Cameletti, 2010).

Decision makers are called to take actions on the basis of the air quality assessment results in order to reduce the impact of air pollution on population health (Scott, 2007). Although uncertainty is crucial to analyze the results, it is not easily interpretable or transmutable into actions on its own, especially when the impact on population health must be evaluated.

The aim of this paper is to provide a statistical framework for the country level population risk assessment connected with air pollution. In particular, the risk is evaluated by estimating the exceedance probability of pollutant concentration thresholds related to a set of airborne pollutants. The exceedance probability is mapped over space and time and, when convolved with the spatial population count distribution, it allows to derive aggregate risk indicator.

The rest of the paper is organized as follows. Section 2 introduces the Dynamic Coregionalization Model which is used to map pollutant concentrations within a multivariate setting and to evaluate uncertainty. Section 3 describes the population risk assessment procedure applied in Section 4 to the Scottish air quality data for the year 2009. Conclusions and future works are reported in Section 5.

2. The dynamic coregionalization model

The Dynamic Coregionalization Model (DCM) is a hierarchical multivariate space-time model introduced by Fassò and Finazzi (2011). Let $y_i(\mathbf{s},t)$ be the concentration of the $i-th$ pollutant at the spatial location $\mathbf{s} \in \mathcal{D} \subset \mathbb{R}^2$ and at time $t \in \mathbb{N}^+$, $1 \leq i \leq q$. The model equation is the following:

$$(1)\quad \begin{bmatrix} y_1(\mathbf{s},t) \\ \vdots \\ y_q(\mathbf{s},t) \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1(\mathbf{s},t) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{x}_q(\mathbf{s},t) \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_q \end{bmatrix} + \mathbf{K}\mathbf{z}(t) + \begin{bmatrix} \delta_1 w_1(\mathbf{s},t) \\ \vdots \\ \delta_q w_q(\mathbf{s},t) \end{bmatrix} + \begin{bmatrix} \varepsilon_1(\mathbf{s},t) \\ \vdots \\ \varepsilon_q(\mathbf{s},t) \end{bmatrix}$$

where $\mathbf{x}_i(\mathbf{s},t)$ and $\boldsymbol{\beta}_i$ are b_i-dimensional vectors of covariates and coefficients respectively. The $p-$dimensional latent state $\mathbf{z}(t)$ has the Markovian dynamics $\mathbf{z}(t) = \mathbf{G}\mathbf{z}(t-1) + \boldsymbol{\eta}(t)$ with $\mathbf{G}$ a stable transition matrix and $\boldsymbol{\eta} \sim N(0, \boldsymbol{\Sigma}_\eta)$. The $q \times p$ matrix $\mathbf{K}$ is the loading matrix of known coefficients. The $\mathbf{w}(\mathbf{s},t) = (w_1(\mathbf{s},t)...w_q(\mathbf{s},t))$ is described by a q-dimensional linear coregionalization model (LCM) of c components $\mathbf{w}(\mathbf{s},t) = \sum_{j=1}^{c} \mathbf{w}^j(\mathbf{s},t)$ where each $w^j(\mathbf{s},t)$ is a latent zero-mean Gaussian process with covariance and cross-covariance matrix function $\boldsymbol{\Gamma}_j = cov\left(w_i^j(\mathbf{s},t), w_{i'}^j(\mathbf{s}',t)\right) = \mathbf{V}_j \rho_j(h, \boldsymbol{\theta}_j)$, $1 \leq i, i^{'} \leq q$, $1 \leq j \leq c$, with $h = \|\mathbf{s} - \mathbf{s}'\|$ the Euclidean distance between s and s' Each $\mathbf{V}_j$ is a correlation matrix and each ρ_j is a valid correlation function. Finally, $\varepsilon_i(\mathbf{s},t) \sim N(0, \sigma^2_{\varepsilon,i})$ is the measurement error which is assumed white-noise in space and time. The model parameter set is $\Psi = \left\{\boldsymbol{\beta}, \boldsymbol{\sigma}^2_\varepsilon; \mathbf{G}, \boldsymbol{\Sigma}_\eta; \boldsymbol{\delta}, \boldsymbol{\theta}, \mathbf{V}\right\}$ where $\boldsymbol{\beta} = \left\{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_q\right\}$, $\boldsymbol{\sigma}^2_\varepsilon = \left\{\sigma^2_{\varepsilon,1}, ..., \sigma^2_{\varepsilon,q}\right\}$, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_c\}$ and $\mathbf{V} = \{\mathbf{V}_1, ..., \mathbf{V}_c\}$. The estimation of Ψ is based on the observed data matrix $\mathbf{Y}(\mathcal{S}, \mathcal{T})$, $\mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_q\} = \left\{\mathbf{s}_{1,1}, ..., \mathbf{s}_{1,n_1}, ..., \mathbf{s}_{q,1}, ..., \mathbf{s}_{q,n_q}\right\} \subset \mathcal{D}$, $\mathcal{T} = \{1, ..., T\}$, $N = n_1 + ... + n_q$ and is carried out through the EM algorithm as discussed in Fassò and Cameletti (2009) and Fassò and Finazzi (2011).

3. Population risk assessment

Given a set of q pollutants, the population risk for the block $B \subset \mathcal{D}$ at time t related to the $i-th$ pollutant is defined here as

$$(2)\quad r_i(B,t) = P_{\hat{\Psi}}\left(y_i(B,t) > L_i\right) \cdot d(B)$$

where $y_i(B,t)$ is the average concentration of the $i-th$ pollutant over the block B, L_i is a concentration threshold and $d(\cdot)$ is the spatial population count distribution, which is here assumed to be time-invariant and error-free. If the block B is small if compared to $\mathcal{D}$, then (2) can be approximated by

$$(3)\quad r_i(B,t) = P_{\hat{\Psi}}\left(y_i(\mathbf{s}^*,t) > L_i\right) \cdot d(B)$$

where $y_i(\mathbf{s}^*,t)$ is the pollutant concentration at site $\mathbf{s}^*$, with $\mathbf{s}^*$ the centre of gravity of B.

When (3) is aggregated over space, the following risk indicator can be defined

$$(4)\quad r_i(t) = \sum_{B \in \mathcal{D}} r_i(B,t)$$

If $d(B)$ is available with the desired spatial resolution, the main problem is how to evaluate the probability of exceedance $P_{\hat{\Psi}}\left(y_i(\mathbf{s}^*,t) > L_i\right)$. Now, the concentration of the $i-th$ pollutant at site $\mathbf{s}^*$ and time t can be obtained by means of the plug-in approach as

$$(5) \quad \hat{y}_i\left(\mathbf{s}^*, t\right) = \mathbf{x}_i\left(\mathbf{s}^*, t\right)\hat{\boldsymbol{\beta}}_i + \mathbf{k}_i\mathbf{z}^T(t) + \hat{\delta}_i\mathbf{w}_i^T\left(\mathbf{s}^*, t\right)$$

where $\mathbf{x}_i\left(\mathbf{s}^*, t\right)$ is a vector of covariates, $\mathbf{k}_i$ is the $i-th$ row of matrix $\mathbf{K}$, $\mathbf{z}^T(t) = E\left[\mathbf{z}(t) \mid \mathbf{Y}(\mathcal{S}, \mathcal{T})\right]$ is the output of the Kalman filter, $\mathbf{w}_i^T(\mathbf{s}^*, t) = E\left[\mathbf{w}_i(\mathbf{s}^*, t) \mid \mathbf{Y}(\mathcal{S}, \mathcal{T})\right]$ is the estimated latent spatial component and $\left\{\hat{\boldsymbol{\beta}}_i, \hat{\delta}_i\right\} \subset \hat{\Psi}$. The variance of $\hat{y}_i\left(\mathbf{s}^*, t\right)$ is denoted by $\hat{\sigma}^2_{\hat{\Psi}}\left(\mathbf{s}^*, t\right)$.

Although the distribution of $\hat{y}_i\left(\mathbf{s}^*, t\right)$ can be assessed for each $\mathbf{s}^* \in \mathcal{D}$ and $t \in \mathcal{T}$, the following inequality holds in general:

$$(6) \quad P_{\hat{\Psi}}\left(y_i\left(\mathbf{s}^*, t\right) > L_i\right) \neq P_{\hat{\Psi}}\left(\hat{y}_i\left(\mathbf{s}^*, t\right) > L_i\right)$$

In order to estimate the lhs of (6) the following procedure is considered:

1. Considering the model in (1) and the dataset $\mathbf{Y}(\mathcal{S}, \mathcal{T})$, the estimate $\hat{\Psi}$ of the model parameter set is provided;

2. Let $Y(\mathcal{S}_{(-j)}, \mathcal{T})$ be the data matrix with the $j-th$ row removed and $\mathbf{y}(\mathbf{s}_j, \mathcal{T})$ the removed row. The leave-one-out cross-validation technique is applied for each $1 \leq j \leq N$ by estimating $\hat{\Psi}^{(-j)}$from $Y(\mathcal{S}_{(-j)}, \mathcal{T})$. The cross-validation residuals are obtained as $e_{\hat{\Psi}}(\mathbf{s}, \mathcal{T}) = \hat{\mathbf{y}}(\mathbf{s}_j, \mathcal{T}) - \mathbf{y}(\mathbf{s}_j, \mathcal{T})$ where $\hat{\mathbf{y}}(\mathbf{s}_j, \mathcal{T})$ is given by (5) for each $t \in \mathcal{T}$.

3. The cross-validation residuals related to the $i-th$ variable are studentized with respect to the dynamic kriging variance $\hat{\sigma}^2_{\hat{\Psi}}(\mathbf{s}, t)$, namely:

$$(7) \quad \tilde{e}_{\hat{\Psi}}(\mathbf{s}, t) = \frac{e_{\hat{\Psi}}(\mathbf{s}, t)}{\hat{\sigma}_{\hat{\Psi}}(\mathbf{s}, t)}; \mathbf{s} \in \mathcal{S}_i, t \in \mathcal{T};$$

4. The cumulative density function $F_{\hat{\Psi}, \tilde{\mathcal{E}}}$ of all the studentized residuals $\tilde{\mathcal{E}} = \left\{\tilde{e}_{\hat{\Psi}}(\mathbf{s}, t) : \mathbf{s} \in \mathcal{S}_i,\ t \in \mathcal{T}\right\}$ is obtained by kernel-smoothing.

5. For each block B and time t, the exceedance probability is evaluated as

$$(8) \quad P_{\hat{\Psi}}(y_i(B, t) > L_i) \equiv 1 - F_{\hat{\Psi}, \tilde{\mathcal{E}}}\left(\frac{L_i - \hat{y}_{i, \hat{\Psi}}(\mathbf{s}^*, t)}{\hat{\sigma}_{\hat{\Psi}}(\mathbf{s}^*, t)}\right)$$

   with $\hat{y}_{i, \hat{\Psi}}(\mathbf{s}^*, t)$ the kriged pollutant concentration under the estimated model with parameter set $\hat{\Psi}$.

Note that (7) is a studentization (Cook, 1982) in a broad sense since $\hat{\sigma}_{\hat{\Psi}}(\mathbf{s}, t)$ is not an estimate of the standard deviation of $e_{\hat{\Psi}}(\mathbf{s}, t)$. However, $\hat{\sigma}_{\hat{\Psi}}(\mathbf{s}, t) \propto \sigma_e(\mathbf{s}, t)$ with $\sigma_e(\mathbf{s}, t)$ the standard deviation of $e_{\hat{\Psi}}(\mathbf{s}, t)$. The studentization is applied in order to homogenize the cross-validation residuals $\tilde{\mathcal{E}}$ which are characterized by heteroscedasticity. Indeed, the exceedance probability in (8) is correctly evaluated provided that the residuals $\tilde{\mathcal{E}}$ are white-noise in space and time.

	$\hat{\beta}_{pop}$	$\hat{\beta}_{slp}$	$\hat{\beta}_{t}$	$\hat{\beta}_{sh}$	$\hat{\beta}_{ws}$	$\hat{\beta}_{blh}$	$\hat{\sigma}^2_{\varepsilon}$
NO_2	0.464	−0.040	0.321	−0.473	−0.197	−0.220	0.367
O_3	−0.090	−0.229	0.392	−0.297	0.216	0.166	0.274
PM_{10}	0.121	0.224	0.289	−0.294	−0.084	−0.222	0.286

Table 1: Estimated β parameters and measurement error variance

4. Scottish air quality data

As an application, a subset of the Scottish air quality data for the year 2009 is considered. In particular, the daily average concentrations of nitrogen dioxide (NO_2), ozone (O_3) and particulate matter (PM_{10}) are jointly modeled by means of (1) and the daily risk indicator (4) is evaluated.

The pollutant concentrations are measured and provided by the Scottish Automatic Urban Network. The number of monitoring sites is 66 for NO_2, 10 for O_3 and 60 for PM_{10}. The population count distribution is provided by the Oak Ridge National Laboratory in the form of the LandScan^{TM} dataset (Bhaduri et al. 2007) and is available with the spatial resolution of 30” × 30” (nearly $1km \times 1km$). The population count is also part of the covariate set which includes the meteorological covariates temperature, sea level pressure, humidity, wind speed and boundary layer height.

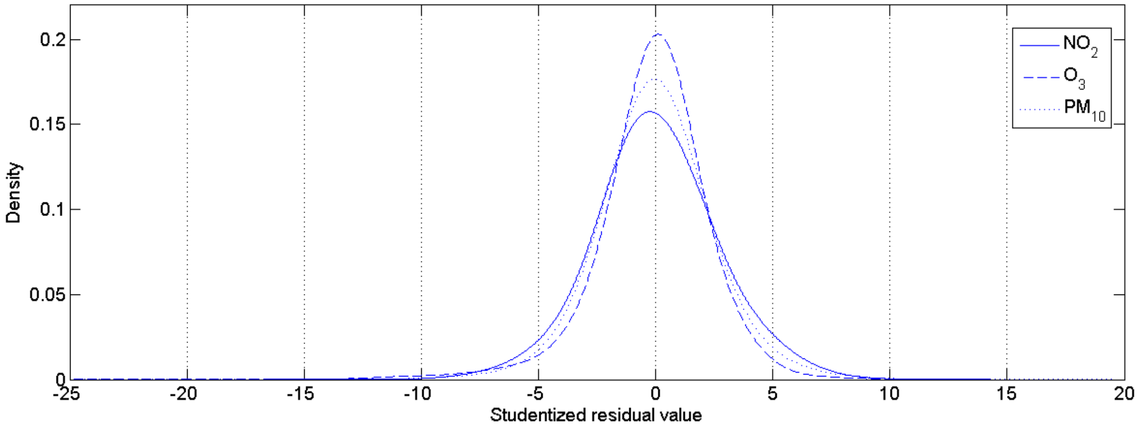

Figure 1: *Studentized residual kernel densities*

The result of the EM algorithm is reported in Table 1 as far as the covariate coefficients and the measurement error variance concern while the rest of the estimated parameters are

$$\hat{\mathbf{G}} = \begin{bmatrix} 0.97 & -0.02 & -0.01 \\ -0.17 & 0.87 & 0.11 \\ 0.27 & 0.13 & 0.58 \end{bmatrix}; \quad \hat{\mathbf{\Sigma}}_{\eta} = \begin{bmatrix} 0.006 & 0.013 & 0.011 \\ 0.013 & 0.063 & -0.026 \\ 0.011 & -0.026 & 0.170 \end{bmatrix}$$

$$\hat{\theta}_1 = 40.99$$

$$\hat{\mathbf{V}}_1 = \begin{bmatrix} 1 & -0.79 & 0.71 \\ -0.79 & 1 & -0.61 \\ 0.71 & -0.61 & 1 \end{bmatrix}$$

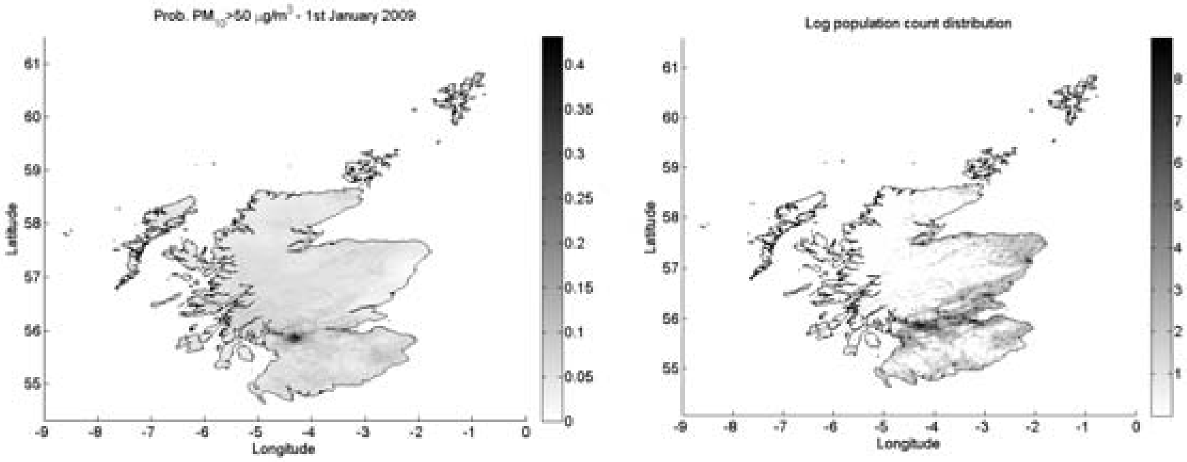

Figure 2: PM_{10} exceedance probability map for the first day of 2009 (lef) and log population count distribution (right).

Note that $p = 3$ and $c = 1$ have been considered and that $\hat{\theta}_1$ is the estimated parameter of the exponential correlation function $\rho\left(h, \boldsymbol{\theta}\right) = exp(-h/\theta)$, $\theta \in \mathbb{R}^+$. The kernel smoothing densities of the studentized residuals related to each pollutant are depicted in Figure (1).

The studentized residuals are used as basis to estimate the exceedance probability as discussed in the previous section. As an example, the PM_{10} exceedance probability map with respect to the concentration threshold $L_{PM_{10}} = 50\,\mu\mathrm{g\,m}^{-3}$ for the first day of 2009 is showed on the left side of Figure (2). The daily exceedance probability maps are used to evaluate the risk as defined in (3) and the risk indicator (4). The daily risk indicator for PM_{10} is depicted in Figure (3). The risk indicator can be directly interpreted as the number of people at risk with respect to the specific pollutant.

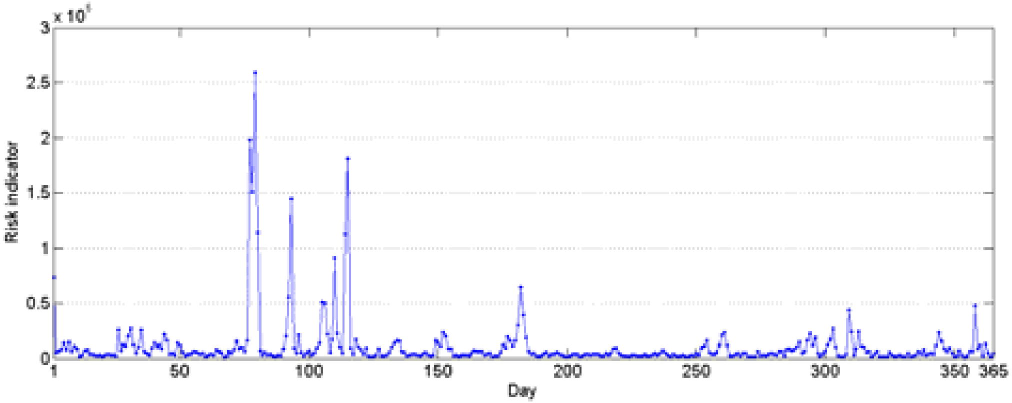

Figure 3: Scottish PM_{10} daily risk indicator for the year 2009 based on the threshold concentration $L_{PM_{10}} = 50\,\mu\mathrm{g\,m}^{-3}$.

5. Conclusion and future works

The Dynamic Coregionalization Model has been considered in order to jointly model and map the pollutant concentration of three main airborne pollutants over Scotland. An approach based on the studentization of the cross-validation residuals has been introduced as a way to estimate the exceedance probability of a pollutant concentration threshold. The exceedance probability is used as basis to estimate the population risk over space and time and an aggregate risk indicator has been provided. In particular, the indicator has been used to evaluate the daily population risk with respect to each pollutant.
Future works will consider the definition of a multi-pollutant risk indicator and the evaluation of confidence intervals on both the exceedance probability and the risk indicator.

Acknowledgements

- This research is part of Project EN17, "Methods for the integration of different renewable energy sources and impact monitoring with satellite data", funded by Lombardy Region under "Frame Agreement 2009".

- The results was made utilizing the LandScan 2008 ^{TM} High Resolution global Population Data Set copyrighted by UT-Battelle, LLC, operator of Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the United States Department of Energy. The United States Government has certain rights in this Data Set. Neither UT-BATTELLE, LLC NOR THE UNITED STATES DEPARTMENT OF ENERGY, NOR ANY OF THEIR EMPLOYEES, MAKES ANY WARRANTY, EXPRESS OR IMPLIED, OR ASSUMES ANY LEGAL LIABILITY OR RESPONSIBILITY FOR THE ACCURACY, COMPLETENESS, OR USEFULNESS OF THE DATA SET.

References

[1] Bhaduri B., Bright E., Coleman P. and Urban M. 2007. LandScan USA: A High Resolution Geospatial and Temporal Modeling Approach for Population Distribution and Dynamics. GeoJournal **69**: 103-117.

[2] Cook R.D. 1982 Residuals and Influence in Regression. New York - Chapman and Hall.

[3] Fassò A, Cameletti M. 2009. The EM algorithm in a distributed computing environment for modelling environmental space-time data. Environmental Modelling & Software **24**: 1027-1035.

[4] Fassò A, Cameletti M. 2010. A unified statistical approach for simulation, modelling, analysis and mapping of environmental data. Simulation. **86**: 139–154.

[5] Fassò A, Finazzi F. 2011. Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. Environmetrics. In printing.

[6] Scott E. M. 2007. Setting and evaluating the effectiveness of environmental policy. Environmetrics **18**: 333-343.