Bayesian spatial SEM for lichen biodiversity

T. Di Battista [1], F. Fortuna [1] P. Giordani [2] and P. Valentini [3]

[1] *Department of Philosophical, Pedagogical and Quantitative Economic Sciences, University G. D'Annunzio, Pescara, Italy; dibattis@unich.it, francesca.fortuna@unich.it*
[2] *DIP.TE.RIS., University of Genoa, Italy; giordani@dipteris.unige.it*
[*] *Department of Economics, University G. D'Annunzio, Pescara, Italy; pvalent@unich.it*

Abstract. *Lichen functional traits are widely used as ecological indicators of environmental quality and, especially, of air pollution. However, due to the high variability of lichen flora across geographic, climatic and ecological gradients, it is often difficult to differentiate the effects of atmospheric pollutants and those of other environmental variables. In order to evaluate this complex relationship, a Bayesian generalized common spatial factor model has been applied to a real data set concerning lichen biodiversity in Liguria region (NW Italy).*

Keywords. *Bayesian spatial SEM; Lichen biodiversity; Air pollution; Biomonitoring techniques; Ecological functional traits.*

---

1 Introduction

The loss of lichen biodiversity in response to environmental conditions is widely used as an indicator of air pollution [1, 3]. However, the high variability of lichen diversity makes difficult to differentiate the effects of atmospheric pollutants and those of other environmental variables since lichen flora highly vary across geographic, climatic and ecological gradients [5].

For this reason, approaches based on biological and ecological species traits (e.g. photosynthetic strategy, growth form, reproductive strategy) have been recently used to assess monitoring change in ecosystems [5, 6]. Basically, functional traits of species are expected to directly link to environmental factors independently of species richness and composition [6] and this allows comparisons among different ecosystems and across regions. However, the influence of environmental conditions on lichen functional traits is still poorly documented, hindering their use in environmental monitoring [5].

In this paper, we thus approach the analysis of multivariate spatial processes from the perspective of recent developments of spatial factor models. Through a fully Bayesian approach, we contribute to the recent empirical literature by melding together factor models, spatial regression models and geostatistical techniques, in order to explain the multifactorial nature of environmental variables and lichen flora. We assume that the relationships existing between the groups of dependent and regressor variables can be studied through a spatially descriptive model, hereafter referred to as spatial structural equation model.

2 The data

Epiphytic lichen biodiversity of Liguria region, in northwestern Italy, has been considered. Data on lichen abundance were collected following the standards suggested by [2]; the survey lasted from 2002 to 2003 and involves a total of 165 sampling sites and a total of 196 epiphytic lichen species [4].
Twelve lichen functional groups have been identified by aggregating the species according to their biological and ecological characteristics [5]. In particular, we have considered the growth form, distinguishing between macrolichens (macro) including both foliose or fruticose species and microlichens (micro); the reproductive strategy of the lichen species: sexual (sex) or asexual (asex) and the nitrogen-tolerance, distinguish between oligotrophic species very sensitive to nitrogen compounds (oligo), mesotrophic requiring intermediate levels of nutrients (meso) and nitrophytic species (nitro) tolerating relatively high level of nitrogen.
Data on atmospheric pollutants were obtained from the Regional Inventory (Liguria Regional Council, 1999). Spot, dispersed and linear emissions of main pollutants were estimated by means of an ISC3 (Industrial Source Complex) long term diffusional model (EPA, 2003). The model has been applied for each $1km^2$ cell on the basis of pollutant concentrations measured by automatic gauges throughout the survey area. For each sampling site, the total annual emissions of the main atmospheric pollutants (NO_x, SO_2 and PM_{10}) in 1999 and 2001 have been calculated as the average of all $1km^2$ cells within a $3km$ buffer.
As potential drivers of lichen diversity [4] three main habitats have been also considered in the model, including broad-leaved forest areas, montane conifers and beech forests and non-forested rural and urban areas.

3 Model

Let $\mathbf{Y}(\mathbf{s}_i)$ and $\mathbf{X}(\mathbf{s}_i)$ be two multivariate spatial processes representing lichen abundance (i.e. the response variable) and pollutants, respectively. We assume that the variables are measured at n sites, $\mathbf{s}_i \in S, i = 1,\dots,n$, with S a spatial domain in two dimensional Euclidean space $\mathcal{R}^2$.
Let $\mathbf{X}_t(\mathbf{s}) = \left(X_{t1}(\mathbf{s}), X_{t2}(\mathbf{s}), X_{t3}(\mathbf{s})\right)'$ be a multivariate spatial process observed at a site $\mathbf{s}$. This process refers to the set of pollutant variables described in section 2 observed at time $t = 1999, 2001$. Since they are defined on positive real line and the standard deviation of response increasing with the mean of the distribution we could model the pollutant data through a gamma distribution as $X_{tj}(\mathbf{s})|\mu_{tj}(s), \alpha_{tj} \sim Ga\left(\alpha_{tj}, \frac{\alpha_{tj}}{\mu_{tj}(s)}\right)$, $j = 1,2,3$, with mean $\mu_{tj}(s)$ and variance $\frac{\mu_{tj}(s)^2}{\alpha_{tj}}$, where α_{tj} is a shape parameter.
The logarithmic link function is used, i.e. $\log(\mu_t(s)) = \nu_{tj}(s) + \lambda_{tj} f_t(s)$, where $\nu_{tj}(s)$ is a general mean, λ_{tj} is the factor loading and $f_t(s)$ is the the spatial common factor at time t. Then, writing $\mathbf{f}_t = \left(f_t(\mathbf{s}_1), \dots, f_t(\mathbf{s}_n)\right)'$, it is assumed that $\mathbf{f}_t$ follows a multivariate normal distribution with mean $\boldsymbol{\mu}_{f_t}$ and a parameterized spatial covariance matrix $\Sigma_f(\phi_t)$ that captures the spatial structure of the common factor, where ϕ_t is the vector of parameters in the covariance structure.
Since the distribution of pollutant factor at $t = 2001$ is strictly related at the polluants at $t = 1999$, as in traditional structural equation models, we relate the two spatial common factors through a linear regres-

sion function. Hence, it is assumed $\mathbf{f}_{2001}|\mathbf{f}_{1999} \sim N\left(\delta\mathbf{f}_{1999}, \Sigma_f(\phi_{2001})\right)$.
Let $\mathbf{Y}^k(\mathbf{s}) = \left(Y_1^k(\mathbf{s}), \ldots, Y_4^k(\mathbf{s})\right)'$ be a multivariate spatial process observed at a site $\mathbf{s}$, where $Y_j^k(\mathbf{s})$ is the process refering to the set of lichen abundance belonging to $k \in K$ where $K = \{Nitro, Meso, Oligo\}$ as described in section 2. Let $Y_j(\mathbf{s})$ be a negative binomial random variable with $E\left(Y_j(\mathbf{s})\right) = \theta_j(\mathbf{s})$ and $Var\left(Y_j(\mathbf{s})\right) = \theta_j(\mathbf{s}) + \kappa_j\theta_j^2(\mathbf{s})$, i.e $Y_j^k(\mathbf{s})|\theta_j^k(\mathbf{s}), \kappa_j^k \sim NB\left(\theta_j^k(\mathbf{s}), \kappa_j^k\right)$. A log link is applied, so that $\log(\theta_j^k(\mathbf{s})) = \log(n(\mathbf{s})) + \psi_j^k g_k(\mathbf{s})$, where $n(\mathbf{s})$ is the number of trees observed at site $\mathbf{s}$, ψ_j^k is the factor loading associated to the latent variable $g_k(\mathbf{s})$. It is worth noting that we define three separate single common factor models based on the cardinality of K, i.e. we assume that the three underlying factors are related only to their own manifest variables. This structure is assumed in order to allow clear interpretation of the latent factors.
Writing $\mathbf{g}_k = \left(g_k(\mathbf{s}_1), \ldots, g_k(\mathbf{s}_n)\right)'$, the structural equation can be defined as

$$\mathbf{g}_k = \mathbf{f}\mathbf{A}_k + \mathbf{Z}\mathbf{B}_k + \mathbf{u}_k \qquad k \in K \tag{1}$$

where $\mathbf{A}_k$ and $\mathbf{B}_k$ are 2×1 and 3×1 coefficient matrices, $\mathbf{f} = [\mathbf{f}_{1999}, \mathbf{f}_{2001}]$, $\mathbf{Z}$ is a matrix of exogenous variables (habitats) and $\mathbf{u}_k$ follows a multivariate normal distribution with zero mean and a parameterized spatial covariance matrix $\Sigma_g(\boldsymbol{\rho}_k)$ that captures the spatial structure of the endogenous common factor, where $\boldsymbol{\rho}_k$ is the vector of parameters in the covariance structure.

4 Inference and computation

4.1 Prior information

The Bayesian specification of the model is completed with the definition of the prior distributions to the model parameters.
In particular, we consider the following priors for the structural equation model: $\mathbf{A}_k \sim N(\mathbf{0}, \Sigma_A)$, $\mathbf{B}_k \sim N(\mathbf{0}, \Sigma_B)$, $\delta \sim N(0, \sigma_\delta^2)$ with $\Sigma_A = \sigma_A^2\mathbf{I}$, $\Sigma_B = \sigma_B^2\mathbf{I}$, and σ_A^2, σ_B^2 and σ_δ^2 large.
For $\mathbf{u}_k$, we choose the exponential correlation function with partial sill ρ_{1k} and decay parameter ρ_{2k} and for $k \in K$. For the decay parameter, we use a discrete uniform prior distribution such that ρ_{2k} can only take values that are within a plausible interval determined by the scale for the data locations. The same strategy is also adopted for the other spatial processes $\mathbf{f}_{1999}$ and $\mathbf{f}_{2001}$.
For all other parameters related to measurement equations, ν_{tj}, λ_{tj} and ψ_j^k, we assume a Normal prior distribution with mean zero and variance chosen to be large to make the priors relatively non-informative. Finally, a non informative Gamma prior is used for the parameters α_{tj} and κ_j^k.

4.2 Posterior inference

Posterior inference for the proposed model is facilitated by MCMC algorithms. Standard MCMC samplers are easily adapted to our model specification such that posterior analysis is readily available. All the parameters are either sampled from normal or gamma full conditional distributions or by simple Metropolis-Hastings steps.

5 Concluding remarks

A distinctive feature of the model is its ability to estimate the factors in the non-normal setting.
In the extended version we will present results from the application.
This model can be extended in a number of directions. The most obvious one is the incorporation of time. Other possible extensions are considering other forms of spatio-temporal data configuration, such as aerial data observed over time.

References

[1] Ammann, K., Herzig, R., Liebendoerfer L. and Urech, M. (1970). Multivariate correlation of deposition data of 8 different air pollutants to lichen data in a small town in Switzerland. *Advanced Aerobiology 51*, 401–406.

[2] Asta, J., Erhardt, W., Ferretti, M., Fornasier, F., Kirschbaum, U., Nimis, P. L., Purvis, O. W., Pirintos, S., Scheidegger, C., Van Haluwyn, C. and Wirth, V. (2002). Mapping lichen diversity as an indicator of environmental quality. In Nimis, P. L., Scheidegger, C., Wolseley, P. A. (eds) Monitoring with Lichens. *Monitoring Lichens*, Nato Science Program-IV, Kluwer Academic Publisher, The Netherlands, 273–279.

[3] Cislaghi, C., Nimis, P. L. (1997). Lichens, air pollution and lung cancer. *Nature* **387**, 463–464.

[4] Giordani, P. (2006). Variables influencing the distribution of epiphytic lichens in heterogeneous areas: a case study for Liguria, NW Italy. *Journal of Vegetation Science* **17**, 195–206.

[5] Giordani, P., Brunialti, G., Bacaro, G., Nascimbene, J. (2012). Functional traits of epiphytic lichens as potential indicators of environmental conditions in forest ecosystems. *Ecological Indicators* **18**, 413–420.

[6] Ricotta, C. and Bacaro, G. (2010). On plot to plot dissimilarity measures based on species functional traits. *Community Ecology* **11**, 113–119.