Montpellier, 13-15 June 2018

Book of Extended Abstracts

Contents

Chapter 1

Contributed oral presentations

Analysing the spatial structure of point patterns and linear networks

Carles Comas[1,*], Sergi Costafreda-Aumedes[4] and Cristina Vega-Garcia [2,3]

[1] *Department of Mathematics, University of Lleida, Agrotecnio Center, Avinguda Estudi General 4, 25001, Lleida, Spain*
[2] *Department of Agricultural and Forest Engineering, Universidad de Lleida, Alcalde Rovira Roure 191, 25198 Lleida, Spain*
[3] *Centro Tecnologico Forestal de Catalunya, Solsona, 25280, Spain*
carles.comas@matematica.udl.cat, cvega@eagrof.udl.cat
[4] *DiSPAA, University of Florence, Piazzale delle Cascine 18, 50144 Florence, Italy; scaumedes@gmail.com*
[*] *Corresponding author*

Abstract. *During the last few decades, it has become increasingly popular the study of events that occur on a network of lines. Examples include, for instance, wildlife-vehicle collisions, street crimes, traffic accidents and plant and tree spatial distribution. For all these cases, as points depend on the linear network, the analysis of such spatial configurations is focused on the description of the spatial configuration of points assuming that the whole point pattern is placed over the linear network. However, in some cases, the dependence between a point pattern and a linear network is not always evident. In these cases, as points do not occur on the linear network, the spatial dependence between point and line segments is not visually obvious (for instance, human-caused fires and road networks). In this work we proposed the definition of a new second order characteristic, based on the Ripley's K function, to analyse the spatial structure between point patterns and linear networks.*

Keywords. *Linear network; Point process; Road network; Spatial point patterns.*

---

1 Introduction

The study of events that occur on a network of lines, such as a road network, has become increasingly popular during the last few decades. Examples include, for instance, wildlife-vehicle collisions [3, 6] and street crimes [1, 5]. In all these examples, point patterns occur on line segments and it is not expected that an event occurs out of these linear networks. As such, the resulting point pattern always depends on the spatial configuration of such linear structures. In these cases, as points depend on the linear network, the analysis of such spatial configurations is focused on the description of the spatial configuration of points assuming that the whole point pattern is placed over the linear network; see for instance [8, 9, 1].

However, in some cases, the dependence between a point pattern and a linear network is not evident. Although the point pattern can depend on such linear structures, points are not constrained to lie along the linear network, but within a certain distance or buffer. In this case, as points do not occur directly over the linear network, the spatial dependence between points and line segments may not be visually apparent. For instance, Human-Caused Fires result in point patterns which spatial structure can depend on the underlying road network [7, 10], but some are placed right on the roadside, and other somewhat

further away, but still benefiting from access provided by the road network.

Nevertheless, few approaches (if any) have been developed to establish a formal approach to describe correlation between point patterns and linear networks. In fact, most of the studies of point patterns and linear networks presuppose that points are constrained to lie along the line segments, thus points locations are determined by line segments. Therefore, our main aim in this work is to propose a new second order measure to analyse the spatial structure between point patterns and linear networks.

2 Linear networks and point processes

Following Ang et al. [1], we define a line segment in the plane with endpoints u and v as $[u,v] = \{tu+(1-t)v : 0 \leq t \leq 1\}$. A linear network L is the union $L = \bigcup_{i=1}^{n_l} l_i$ of a finite collection of line segments $l_1,\ldots,l_{n_l}$ in the plane, with a total length $|L|$. We generalize the definition of L and we also assume that L represents a stochastic mechanism on a region $A \subset \mathbb{R}^2$ that can generate also such linear networks. Moreover, we consider that this stochastic mechanism L is stationary and isotropic, in the sense that the resulting statistical properties of L are the same in different, but geometrically similar, subregions of A; *isotropy* means that there are no directional effects. Moreover, we consider a spatial stationary point process as a stochastic mechanism Φ that generates a countable set of events $\mathbf{x}_i$ in a bounded region A, with point intensity λ (see, for instance, [4]). We do not presuppose Φ to be initially related to L. Our intention is to obtain a second-order measure to detect spatial structures between point patterns and linear Networks. With this in mind, we propose an extension of the Ripley's K-function to detect such structures. Let define a second-order characteristic, based on the Ripley's K-function, to analyse the spatial structure between Φ and a linear network L via

$$K_{LX}(r) = \frac{1}{\lambda_L}\mathbb{E}\Big[\int_L \mathbf{I}(0 < \|\mathbf{x}-y\| \leq r)dy|\mathbf{x} \in \Phi\Big] \tag{1}$$

where $\|\mathbf{x}-y\|$, the Euclidean norm, computes the Euclidean distance between a point $\mathbf{x} \in \Phi$ and a point y located in L, $\lambda_L = |L|/|A|$ is the intensity of L and $\mathbf{I}(x)$ is the indicator function, where $\mathbf{I}(x) = 1$ if x is true and $\mathbf{I}(x) = 0$ otherwise. Then $\lambda_L K_{LX}(s)$ is the expected length of L falling in a disk $b(\mathbf{x},r)$ with radius r and centrered at $\mathbf{x} \in \Phi$. It is straightforward to deduce that the expected length of L falling in a disk with radius r located at any random location of A is $\lambda_L \times |b(x,r)|$. Dividing by λ_L shows that in the case that a point process Φ and a linear network L are independent, $K_{LX}(r) = \pi r^2$.

We propose a non-parametric estimator of this function based on moment estimators and approximating the integral term by a Riemann sum, and thus

$$\begin{aligned}\hat{K}_{LX}(r) &= \frac{|A|}{|L|}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_L}\omega_{ij}^{-1}\mathbf{I}(0 < \|\mathbf{x}_i-\mathbf{y}_j\| \leq r)\frac{|L|}{n_L} \\ &= \frac{|A|}{nn_L}\sum_{i=1}^{n}\sum_{j=1}^{n_L}\omega_{ij}^{-1}\mathbf{I}(0 < \|\mathbf{x}_i-\mathbf{y}_j\| \leq r)\end{aligned} \tag{2}$$

where n_L is the number of partitions of L for the Riemann sum, $\|\mathbf{x}_i-\mathbf{y}_j\|$ is the Euclidean distance between points of the point pattern ϕ and middle points of the resulting line segments of the n_L partitions, n is the number of points of the point pattern ϕ, and ω_{ij} is the Ripley's factor [11] to correct for edge effects. Note that this factors is obtained as assuming that the resulting middle points for the n_L segments are in fact a new point pattern. Also notice that n_L should be large to obtain a good approximation of the integral term in (1).

3 Simulation study

We conducted a simulation study to illustrate the use of the new second-order measure to detect correlation structures between point patterns and linear networks under several point configurations and a linear network defined by a road configuration. This linear network is a real road network from a square region (30 km $\times$ 30 km) in Asturias (North of Spain). We considered several point process mechanisms to generate distinct point configurations in the unit square, assuming this road network. To generate cluster structures between a point pattern and a linear network (i.e. attraction between points and line segments), we considered the Spatstat R package [2] to generate realisations of Poisson point processes with a specified point intensity on a linear network. The resulting point pattern depends on the linear network as all the points are placed on the linear structure. To reduce the degree of dependence between points and line segments, we assumed random shifts of the resulting Poisson point pattern on the linear network. In fact, for $\mathbf{x}_i = (x_i, y_i)$, for $i = 1, \ldots, n$, we took $x_i \to x_i + aU(-1,1)$, and $y_i \to y_i + aU(-1,1)$, where $a \geq 0$ is a constant that defines the strength of attraction between points and line segments; for $a = 0$ the point pattern is not shifted and we obtain the original Poisson pattern on the linear network. Moreover, we also considered realizations of repulsion structures between points and line segments. This point process is essentially defined as a stationary and isotropic Poisson point process over A, where we impose a minimum repulsion distance between points and the linear network. Here immigrants arrive randomly in time according to a Poisson process with rate α and have uniformly distributed locations on A. If the minimum distance between the newly arrived point and L is less than a prescribed repulsion distance d_r then newly arrived points are accepted with probability p. Otherwise, if this minimum distance is larger than or equal to d_r, newly arrived points are always accepted. To obtain (2), we consider $n_L = 299757$ partitions to obtain a reasonable approximation for the integral term.

Figure 1 shows the resulting point patterns for the attraction and repulsion scenarios, for $a = 500$ meters (middle attraction effects), and $d_r = 200$ meters and $p = 0.5$ (middle repulsion effects), respectively, for a point intensity of around 1000 points. This highlights that for the attraction example, points are clumped together in spatial regions where the density of roads are higher, and under the repulsion scenario points form clusters avoiding line segments. Inspection of the resulting empirical $\hat{K}_{LX}(r) - \pi r^2$ function confirms this results. In both cases, this empirical function lies outside the upper and lower envelopes based on 1000 realizations defined as random shifts of the analysed point pattern (random superposition of points and line segments).

4 Future work

The next steps in our work are to extent the simulation study assuming further scenarios (attraction and repulsion) and consider a real data set involving human-caused fires and road networks.

Acknowledgments. Work funded by grant MTM2017-86767-R from the Spanish Ministry of Economy, Industry and Competitiveness

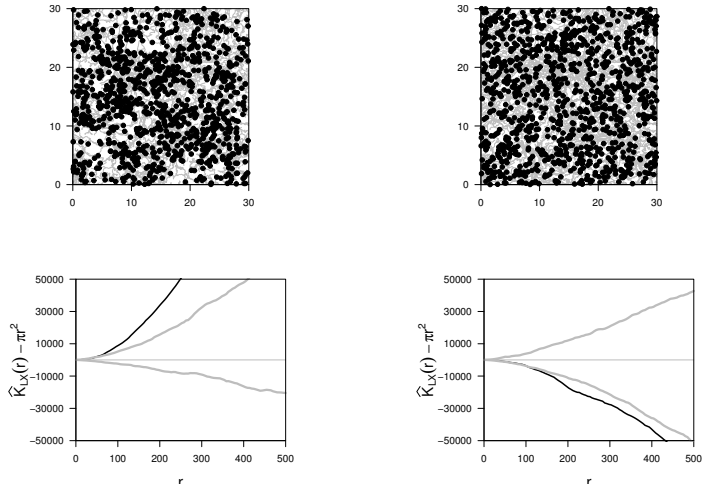


Figure 1: Attraction and inhibitory point patterns between points and a road network (left and right columns, respectively) for $a = 500$ meters (attraction), and $d_r = 200$ and $p = 0.5$ (repulsion), together with resulting empirical functions $\hat{K}_{LX}(r) - \pi r^2$ corresponding to the attraction and repulsion point patterns (solid line) together with the maximum and the minimum envelopes (grey line) based on 1000 realizations defined as random shifts of the analysed point pattern.

References

[1] Ang, W., Baddeley, A., and Nair, G. (2012). Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology . *Scandinavian Journal of Statistics*. **39**, 591–617.

[2] Baddeley, A., Rubak, E. and Turner, R. (2015). Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC Press. London.

[3] Diaz-Varela, E. R., Vazquez-Gonzalez I, and Marey-Perez, M.F. (2011). Assessing methods of mitigating wildlife-vehicle collisions by accident characterization and spatial analysis. *Transport Research Part D* **16**, 281–287.

[4] Diggle, P. J. (2003). Statistical Analysis of Spatial Point Patterns. Hodder Arnold.

[5] Eckardt, M. and Mateu, J. (2017). Analysing highly complex and highly structured point patterns in space. *Spatial Statistics (In press)*.

[6] Morelle, K., Lehaire, F. and Lejeun, P. (2013). Spatio-temporal patterns of wildlife-vehicle collisions in a region with a high-density road network. *Nature Conservatio*. **5**, 53–73.

[7] Morrison, P. H. (2007). Roads and Wildfires. Pacific Biodiversity Institute. (Winthrop, Washington).

[8] OKabe, A. and Yamada, I. (2001). The K-function method on a network and its computational implementation. *Geografical Analysis*. **33**, 271–290.

[9] OKabe, A., Okunuki, K., and Shiode, S. (2006). SANET: a toolbox for spatial analysis on a network. *Geografical Analysis*. **28**, 57–66.

[10] Penman, T. D., Bradstock, R. A. and Price, O. (2013). Modelling the determinants of ignition in the Sydney Basin, Australia: implications for future management. *International Journal of Wildland Fire*. **22**, 469–478.

[11] Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability*. **13**, 255–266.

Spatio-temporal point processes: a second-order marked framework

F. Ballani[1], F. J. Rodríguez-Cortés[2,*], J. Mateu[2] and D. Stoyan[1]

[1] *Institut für Stochastik, TU Bergakademie Freiberg, Freiberg, Germany; ballani@math.tu-freiberg.de, stoyan@math.tu-freiberg.de*
[2] *Department of Mathematics, University Jaume I, Castellón, Spain; cortesf@uji.es, mateu@uji.es*
[*] *Corresponding author*

Abstract. *Spatio-temporal point processes can be analysed statistically by considering the times as marks of the locations or the locations as marks of the times. For these marked point processes, classical second-order characteristics yield interesting information about spatio-temporal correlations. These summary functions provide valuable information about the relations between the points over distances in space and time and on the strength and range of interaction in spatio-temporal patterns. We give statistical estimators of these summary characteristics and investigate their properties by simulation.*

Keywords. *Mark summary statistics; Marking; Second-order characteristics; Spatio-temporal point patterns.*

---

1 General set-up

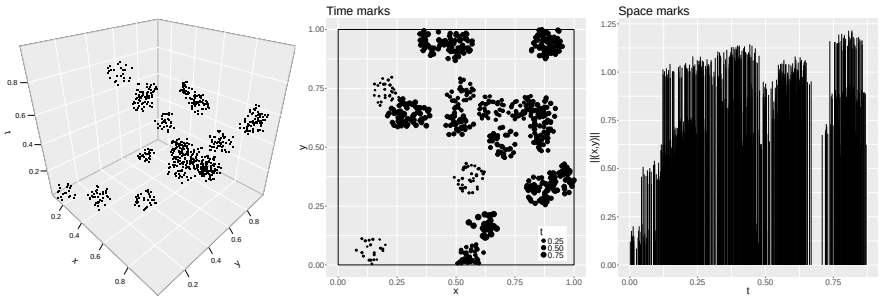
Paper [3], which in the following is denoted as Part 1, suggested the idea of transforming a spatio-temporal point process into marked point processes, and of analysing them by means of known second-order summary functions of marked point processes. We can consider the times as marks of the spatial point process of point locations, which yields a spatial point process with real-valued marks, or we can consider the locations as marks of the times, which yields a one-dimensional point process on the time axis with vector-valued marks. While marks for usual point processes are often considered only an additional issue, for spatio-temporal processes they are an inherent, quite natural and a-priori given issue.

The definitions and notation introduced in Part 1 are used also here. We consider a spatio-temporal point process N in $\mathbb{R}^d \times \mathbb{R}$ as a random sequence of points

$$N = \{[x_1,t_1],[x_2,t_2],\ldots\},$$

where the x_n are spatial locations and t_n the corresponding times. We assume that the point process N is orderly, roughly meaning that coincident points do not occur. In the theoretical part of the paper we assume that the basic point process N is *completely stationary* and in the context of summary characteristics we additionally assume *complete isotropy*, which means that all rotations of N around the origin of $\mathbb{R}^d \times \mathbb{R}$ have the same distribution as the original spatio-temporal point process N. A completely stationary spatio-temporal point process has a constant *intensity* λ, defined as $\lambda = \mathbf{E}(N(\boxed{1} \times \boxed{1}))$, i. e., λ is the

mean number of points per unit volume and unit time. The first $\boxed{1}$ denotes the unit cube and the second the unit interval.



(a) Spatio-temporal pattern (b) Time marks (c) Location marks

Figure 1: Realisation of a spatio-temporal double-cluster point process in $\boxed{1}\times\boxed{1}$ with parameters $\lambda_p = 20$, $R = 0.07$ and $\mu = 100$. We additionally represent the corresponding time and location marks, where the location mark is given by the norm.

A simulated realisation of a double-cluster point process on the spatio-temporal window $W \times T = \boxed{1}\times\boxed{1}$ is shown in Figure 1, together with the corresponding time and location marks, where the spatial mark is given by the norm of the location vector. Again, in the time mark plot the diameters of the circles are proportional to the temporal unit, with sizes of the points indicating older points.

2 Mark summary functions for spatio-temporal point processes

2.1 Time marks (M_T)

For the real-valued time marks of M_T the theory in [1] can be applied without changes. We have only to note that the mean mark $\overline{m}_{\text{sp}}$ is $\tau/2$. A *test function* is a measurable function $f\colon \mathbb{R}^2 \to [0,\infty)$ that assigns a non-negative number to two marks. The corresponding (non-normalised) mark correlation function $c_{f,\text{sp}}(r)$ can be written as

$$c_{f,\text{sp}}(r) = \mathbf{E}\left[f(t(x_1),t(x_2))\,|\,x_1,x_2 \in N_{\text{space}}, \|x_1 - x_2\| = r\right], \quad r > 0. \tag{1}$$

Here N_{space} is the spatial component of the spatio-temporal point process, and $t(x_1)$ and $t(x_2)$ are the time marks of the points x_1 and x_2. The function $c_{f,\text{sp}}(r)$ can be expressed (and estimated) by means of the product densities $\rho_{\text{sp}}(r)$ and $\rho_{f,\text{sp}}(r)$, where $\rho_{\text{sp}}(r)$ is the second-order product density of N_{space} and $\rho_{f,\text{sp}}(r)$ the density with respect to the $2d$-dimensional Lebesgue measure of the second-order factorial measure $\alpha_{f,\text{sp}}$. We then have

$$c_{f,\text{sp}}(r) = \frac{\rho_{f,\text{sp}}(r)}{\rho_{\text{sp}}(r)}, \quad r > 0.$$

The mark variogram $\gamma_{\text{sp}}(r)$ [1], which is a non-normalised characteristic, is obtained by

$$\gamma_{\text{sp}}(r) = c_{v,\text{sp}}(r) - c_{c,\text{sp}}(r), \quad r > 0.$$

For more details see [2]. The most frequently used *normalised* correlation functions are the *mark correlation function* $k_{mm,\mathrm{sp}}$,

$$k_{mm,\mathrm{sp}}(r) = c_{c,\mathrm{sp}}(r)/\overline{m}^2_{\mathrm{sp}} = c_{c,\mathrm{sp}}(r)/(\tau/2)^2, \quad r > 0.$$

2.2 Location marks (M_W)

A test function is now a measurable function $f\colon \mathbb{R}^{2d} \to [0,\infty)$ that assigns a non-negative number to two location marks. The corresponding (non-normalised) mark correlation function $c_{f,\mathrm{te}}(t)$ can be written as

$$c_{f,\mathrm{te}}(t) = \mathbf{E}\left[f(x(t_1),x(t_2)) \,|\, t_1,t_2 \in N_{\mathrm{time}}, |t_2 - t_1| = t\right], \quad t > 0. \tag{2}$$

Here N_{time} is the temporal component of the spatio-temporal point process, and $x(t_1)$ and $x(t_2)$ are the location marks of the time points t_1 and t_2. Temporal counterparts of the mark summary functions can be defined using the test functions in the same fashion as in previous case, i. e., we consider in what follows a corresponding

- *norm-mark variogram* $\gamma^{[n]}_{\mathrm{te}} = c_{v,\mathrm{te}} - c_{c,\mathrm{te}}$,
- the *norm-mark correlation function* $k^{[n]}_{mm,\mathrm{te}} = c_{c,\mathrm{te}}/\overline{m}^2_{\mathrm{te}}$.

3 Non-parametric estimators for the mark summary functions

We recommend the statistical estimators of Part 1 also here and thus we have the following estimators. The estimator of $c_{f,\mathrm{sp}}(r)$ is given by

$$\widehat{c}_{f,\mathrm{sp}}(r) = \frac{\sum^{\neq}_{x_1,x_2 \in N_{\mathrm{space}} \cap W} f(t_1,t_2)\kappa_\varepsilon(\|x_1 - x_2\| - r)}{\sum^{\neq}_{x_1,x_2 \in N_{\mathrm{space}} \cap W} \kappa_\varepsilon(\|x_1 - x_2\| - r)}, \quad r > \varepsilon > 0,$$

and that of $c_{f,\mathrm{te}}(t)$ by

$$\widehat{c}_{f,\mathrm{te}}(t) = \frac{\sum^{\neq}_{t_1,t_2 \in N_{\mathrm{time}} \cap T} f(x_1,x_2)\kappa_\delta(|t_2 - t_1| - t)}{\sum^{\neq}_{t_1,t_2 \in N_{\mathrm{time}} \cap T} \kappa_\delta(|t_2 - t_1| - t)}, \quad t > \delta > 0,$$

where κ_ε and κ_δ are one-dimensional kernel functions with spatial bandwidth ε and temporal bandwidth δ, respectively.

4 Simulation study

Figure 2 shows the estimated and theoretical spatial mark variogram $\gamma_{\mathrm{sp}}(r)$ for the double-cluster model indicating spatio-temporal interaction up to distances from 0.13 and 0.15 which are close to the spheres

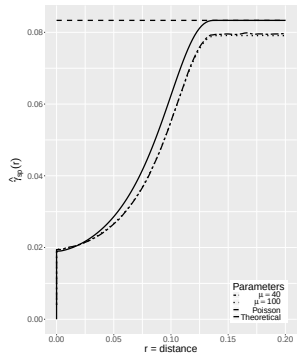
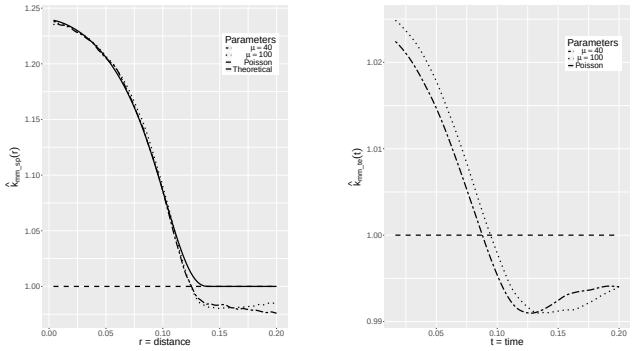


Figure 2: Estimated and theoretical mark variograms $\gamma_{\mathrm{sp}}(r)$ for a double-cluster model together with the constant value for the Poisson case. The bandwidths are ($\mu = 40$: $\varepsilon = 0.003$) and ($\mu = 100$: $\varepsilon = 0.024$).



(a) Spatial mark correlation (b) Temporal mark correlation

Figure 3: Estimated and theoretical mark correlation $k_{mm,\mathrm{sp}}(r)$ (left) and estimated $k_{mm,\mathrm{te}}(t)$ (right) for a double-cluster model together with the constant value for the Poisson case.

diameter. Further, these values are also close to the range $2R = 0.14$ with sill $1/12 = 0.083$ for the theoretical spatial mark variogram. The nugget effect 0.019 suggests a high variability of locations for small distances.

Figure 3a shows the theoretical and estimated $k_{mm,\mathrm{sp}}(r)$ together with the constant value for the Poisson case. In both cases, the theoretical and estimated spatial mark correlation functions are close together and follow the same behaviour over the range of distances. This means that the conditional mean of the product of marks given that there is a pair of points in the unmarked pattern with inter-event distances approximately equal to r or t, are relatively more frequent compared to the case of a Poisson process.

References

[1] Illian, J., Penttinen, A., Stoyan, H., Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley & Sons, Chichester.

[2] Schlather, M. (2001). On the second-order characteristics of marked point processes. *Bernoulli*, **7**: 99–117.

[3] Stoyan, D., Rodríguez-Cortés, F. J., Mateu, J. and Gille, W. (2017). Mark variograms for spatio-temporal point processes, *Spatial Statistics*, **20**: 125–147.

Testing in spatial nonhomogeneous Poisson point processes with covariates

M.I. Borrajo[1,*], W. González-Manteiga[2] and M.D. Martínez-Miranda[3]

[1] *Department of Statistics and Operations Research and Mathematical Didactics, University of Oviedo; maribelborrajo@uniovi.es*
[2] *Department of Statistics, Mathematical Analysis and Optimisation, University of Santiago de Compostela*
[3] *Department of Statistics and Operations Research, University of Granada*
[*] *Corresponding author*

Abstract. *The first-order intensity function is one of the functions characterising a point process, and its study has been approached so far from different perspectives. One appealing model describes the intensity as a function of a spatial covariate, and in the recent literature estimation theory and several applications have been developed under this model. In this work we first formulate a goodness-of-fit test for this intensity model, assuming a nonhomogeneous Poisson point processes, and secondly we formulate a two sample testing problem. Both tests are based on an L^2-distance, their normal asymptotic distributions are proved and appropriate bootstrap procedures are implemented to calibrate them. The performance of the proposed techniques is analysed in extensive simulation studies and is illustrated with two real data sets: Murchison gold deposits in Western Australia and wildfire data in Canada.*

Keywords. *Point processes; First-order intensity; Testing; Covariates; Wildfires.*

---

1 Introduction

Analysing the dependence of a point process on spatial covariates has generated an increasing interest in the last decades. The possible applications in fields such as ecology, forestry, seismology and epidemiology among others have been on the basis of this raise of attention to the problem.

First-order intensity function is one of the characteristic functions of a point process and its study is one of the main aims in this field. Assuming a parametric model may be a way of estimating it, using for instance a likelihood score such as the Akaike information criteria (AIC), see [6] or pseudolikelihood procedures, see for example [11]. In the Bayesian context [10] proposed some models based on log-gaussian Cox processes. However, it is well-known that these techniques can provide unreliable estimations when the assumed model does not fit the real intensity. Hence, an alternative through nonparametric methods such as kernel estimation may apply.

The first kernel intensity estimator was proposed in [5], which has been mostly used in exploratory analysis due to its lack of consistency. In the last decades, this lack of consistency as well as some real applications requirements, induced a new scenario based on the inclusion of covariates in the model. [9]

proposed a kernel intensity estimator, assuming that the intensity function depends on some observed spatially varying covariates through an unknown continuous function. Later, [1] postulate

$$\lambda(x) = \rho(Z(x)), x \in W \subset \mathbb{R}^2, \tag{1}$$

where $Z : W \subset \mathbb{R}^2 \to \mathbb{R}$ is a spatial continuous covariate that is exactly known in every point of the region of interest W. In practice this covariate will commonly be know in an enough amount of points spread over the region, so the values for the rest of the points can be interpolated and it can be assumed that these values are indeed the real ones.

The inclusion of spatially varying covariates has been a big step forward on point process theory; however, a little attention has been paid to test the dependence on these covariates. To the extent of our knowledge, only something similar has been done in a slightly different context in [4], where the authors assume that the spatio-temporal intensity depends on a linear combination of several covariates and they test, using the conditional intensity, whether any of the coefficients are null.

In this work we try to fulfil the existing gap on checking the goodness-of-fit of model (1) by defining a suitable testing procedure. Moreover, once assumed the hypothesis of the covariate dependence as it is indicated in (1), we formulate a two sample problem where we test if the first-order intensities of two spatial point patterns are equal through their densities of event locations.

2 Goodness-of-fit test

Let X be a point process defined in a region $W \subset \mathbb{R}^2$, where W is assumed to have finite positive area; $X_1, \ldots, X_N$ a realisation of the process, where N is the random variable counting the number of events, and $Z : W \subset \mathbb{R}^2 \to \mathbb{R}$ the spatial continuous covariate.

We formulate the null hypothesis $H_0 : \lambda(x) = \rho(Z(x)), x \in W$ versus a general alternative in which the intensity function is not explained completely through the covariate. The idea is to define a test statistic based on a L^2-distance between the classical kernel intensity estimator defined by [5] and the intensity estimator under model (1) proposed by [2]. Due to the lack of consistency of Diggle's proposal we have decide to do a equivalent comparison using the concept of "density of events location" of [3] instead of using the intensities; i.e., the null hypothesis can be equivalently rewritten as $H_0 : \lambda_0(x) = \rho(Z(x))/m$, with $\lambda_0(x) = \lambda(x)/m$ and $m = \int_W \lambda(x)dx$.

Hence, the test statistic is defined as:

$$S = \int_W \left(\hat{\lambda}_{0,H}(x) - \hat{\rho}_{0,b}(Z(x))\right)^2 dx, \tag{2}$$

where $\hat{\lambda}_{0,H}(x) = \frac{1}{N p_H(x)} \sum_{i=1}^{N} \boldsymbol{K}_H(x - X_i) 1_{\{N \neq 0\}}$ is the bivariate estimation for the density of events location proposed by [7], H is a bandwidth matrix, $p_H(x)$ is the edge correction term, $\hat{\rho}_{0,b}(Z(x)) = \frac{\hat{\rho}_b(x)}{N} 1_{\{N \neq 0\}}$ with $\hat{\rho}_b(x) = \sum_{i=1}^{N} \frac{1}{g^*(Z(X_i))} L_b(Z(x) - Z(X_i))$, b is a scalar bandwidth parameter, $\boldsymbol{K}$ and L are kernel functions, $\boldsymbol{K}_H(u) = |H|^{-1/2} \boldsymbol{K}(H^{-1/2}u)$, $|\cdot|$ denotes the determinant of a matrix, $L_b(u) = \frac{1}{b} L\left(\frac{u}{b}\right)$ and g^* is the unnormalised version of the derivative of the cumulative distribution function $G(z) = \int_W 1_{\{Z(u) \leq z\}} du$.

Under some regularity conditions, we determine the asymptotic distribution of the test, and it holds that $\frac{S - \mu_S}{\sigma_S} \longrightarrow N(0,1)$, where $\mu_S = A(m)|H|^{-1/2} R(\boldsymbol{K}) + \frac{1}{2}\boldsymbol{\mu}_2(\boldsymbol{K}) \int \lambda_0(x) tr^2(HD^2\lambda_0(x))dx + \frac{1}{4}\boldsymbol{\mu}_2^2(\boldsymbol{K})$

$\int tr^2(HD^2\lambda_0(x))dx$ and $\sigma_S^2 = A(m)|H|^{-1/2}\int\int\lambda_0^2(x)\lambda_0(y)(\boldsymbol{K}\circ\boldsymbol{K})(H^{-1/2}(x-y))dxdy + 2A(m)|H|^{-1/2}$ $R(\lambda_0)R(\boldsymbol{K})$, with $tr(\cdot)$ denoting the trace of a matrix, $\circ$ the convolution between two functions, D^2 the matrix of the second order derivatives, $A(m) = E[\frac{1}{N}1_{\{N\neq 0\}}]$, $1_{\{\cdot\}}$ the indicator function, $\boldsymbol{\mu}_2(\boldsymbol{K}) = \int_{\mathbb{R}^2} uu^T\boldsymbol{K}$ and $R(\cdot)$ the integral of the square of a function.

However, this asymptotic distribution requires some extra estimations, and as the convergence rate may be slow it is not suitable for small patterns. Our proposal to deal with this inaccuracy is to use a smooth bootstrap procedure to resample under the null hypothesis and calibrate the test.

The performance of all this methodology is analysed in an extensive simulation study including several models with different covariates, in which we study the level as well as the power values of the test. Moreover the proposed test is applied to two real data sets: the one formed by the Murchison gold deposits with the distance to the geological faults as covariate, and the one composed by the wildfires in Canada during June 2015 with meteorological covariates.

3 Two sample problem under the covariate dependent intensity model

Our aim in this section has been briefly pointed out by the end of the introduction: we want to test whether two given independent patterns are originated by the same process, assuming that the theoretical intensity depends on a known covariate in the way shown in (1). To check this hypothesis, we define a new L^2-distance based test statistic.

Let X_i with $i = 1,2$ be two point processes defined in a region $W \subset \mathbb{R}^2$, where W is assumed to have finite positive area, $X_{11},\ldots,X_{1N_1}$ and $X_{21},\ldots,X_{2N_2}$ be two realisations of the processes where N_i are the random variables counting the number of events and recall $Z : W \subset \mathbb{R}^2 \to \mathbb{R}$ the spatial continuous covariate exactly known in every point of the region of interest W. We denote by $Z_{11},\ldots,Z_{1N_1}$ and $Z_{21},\ldots,Z_{2N_2}$ the realisations transformed through the covariate, i.e., $Z_{ij} = Z(X_{ij})$.

Under model (1) let denote by $\lambda_i(x) = \rho_i(Z(x))$ the intensity functions corresponding to the processes X_i with $i = 1,2$. We want to test the null hypothesis $H_0 : \lambda_1(x) = \lambda_2(x), x \in W$ versus the two-sided alternative. Following what it has been done in [3], [2] and [8], we use the density of events location to define an equivalent null hypothesis. Hence, let $f_i(z) = \frac{g^\star(z)\rho_i(z)}{m_i}$; then, $H_0 : f_1(z) = f_2(z),\ \ z \in \mathbb{R}$. Remark that this does not really need to be in $\mathbb{R}$ but in a subset of it covering the range of values of the covariate Z.

To address the problem of defining an appropriate test, we need a measure between the two theoretical densities which we use to define our statistic. In this case we have chosen the L^2-distance and we can define our test statistic $\mathcal{S} = \widehat{\psi}_{11} + \widehat{\psi}_{22} - \widehat{\psi}_{12} - \widehat{\psi}_{21}$, where taking into account the estimator introduced in [2], we can define

$$\widehat{\psi}_1 = \frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\frac{g^\star(Z_{1i})}{g^\star(Z_{1j})}L_{h_1}(Z_{1i}-Z_{1j})1_{\{N_1\neq 0\}},\quad \widehat{\psi}_{12} = \frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\frac{g^\star(Z_{2j})}{g^\star(Z_{1i})}L_{h_1}(Z_{2j}-Z_{1i})1_{\{N_1\neq 0,N_2\neq 0\}},$$

$$\widehat{\psi}_2 = \frac{1}{N_2^2}\sum_{i=1}^{N_2}\sum_{j=1}^{N_2}\frac{g^\star(Z_{2i})}{g^\star(Z_{2j})}L_{h_2}(Z_{2i}-Z_{2j})1_{\{N_2\neq 0\}},\quad \widehat{\psi}_{21} = \frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\frac{g^\star(Z_{1i})}{g^\star(Z_{2j})}L_{h_2}(Z_{1i}-Z_{2j})1_{\{N_1\neq 0,N_2\neq 0\}},$$

with h_i scalar bandwidths and L a univariate kernel function.

Under some regularity conditions, we determine the asymptotic distribution of the test, and it holds that $\frac{S-\mu_S}{\sigma_S} \longrightarrow N(0,1)$, where $\mu_S = (A(m_1)h_1 + A(m_2)h_2)L(0) + o(A(m_1)) + o(A(m_2))$ and $\sigma_S = 2B(m_1)\frac{1}{h_1}R(L)\psi + 2B(m_2)\frac{1}{h_2}R(L)\psi + A(m_1)A(m_2)\psi R(L)\left(\frac{1}{h_1}+\frac{1}{h_2}\right) + A(m_1)A(m_2)\psi\left(\frac{1}{h_1}\int L(u)L_{h_2/h_1}(u)du + \frac{1}{h_2}\int L(u)L_{h_1/h_2}(u)du\right) + O(B(m_1) + O(B(m_2)))$, with $B(m_i) = \mathbb{E}\left[\frac{1}{N_i^2}1_{\{N_i\neq 0\}}\right]$ and $\psi \equiv \psi_{11} \equiv \psi_{22} \equiv \psi_{12} \equiv \psi_{21}$ under the null.

This asymptotic distribution also requires some extra estimations and, the slow convergence rate is present here, son it may not be suitable in all the situations. Hence, we propose again to use an adapted bootstrap procedure to calibrate the test.

The finite sample properties of the two sample test are evaluated through a simulation study. The simulated models have been defined to represent real data sets such as the Murchison data and the Canada wildfires data.

References

[1] Baddeley, A.; Chang, Y. M.; Song, Y. and Turner, R. (2012) Nonparametric estimation of the dependence of a spatial point process on spatial covariates. *Statistics and Its Interface* **5**, 221–236.

[2] Borrajo, M. I.; González-Manteiga, W. and Martínez-Miranda, M. D. (Under review) Bootrapping kernel Intensity estimation for nonhomogeneous point processes depending on spatial covariates.

[3] Cucala, L. (2006) Espacements bidimensionnels et données entachés d'erreurs dans l'analyse des procesus ponctuels spatiaux. *Université des Sciences de Toulouse I.*

[4] Díaz-Avalos, C.; Juan, P. and Mateu, J. (2014) Significance tests for covariate-dependent trends in inhomogeneous spatio-temporal point processes. *Stochastic environmental research and risk assessment* **28**, 593–609.

[5] Diggle, P. J. (1985) A kernel method for smoothing point process data. *Journal of the Royal Statistical Society. Series C* **34**, 138–147.

[6] Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. CRC Press.

[7] Fuentes-Santos, I.; González-Manteiga, W. and Mateu, J. (2015) Consistent smooth bootstrap kernel intensity estimation for inhomogeneous spatial Poisson point processes. *Scand. Journal of Statistics* **43**, 416–435.

[8] Fuentes-Santos, I.; González-Manteiga, W. and Mateu, J. (2017) A nonparametric test for the comparison of first-order structures of spatial point processes. *Spatial Statistics* **22**, 240–260.

[9] Guan, Y. (2008). On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. *Journal of the American Statistical Association* **103**, 1238–1247.

[10] Illian, J.; Sørbye, S. H. and Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *The Annals of Applied Statistics* **6**, 1499–1530.

[11] Waagepetersen, R. P. (2007) An estimating function approach to inference for inhomogeneous Neyman–Scott processes. *Biometrics* **63**, 252–258.

Detecting and modeling multi-scale space-time structures of wildfire occurrences

Edith Gabriel[1], Thomas Opitz[2] and Florent Bonneu[1,*]

[1] *LMA EA2151, Avignon University, F-84000 Avignon, France; edith.gabriel@univ-avignon.fr, florent.bonneu@univ-avignon.fr*
[2] *BioSP, INRA, F-84914 Avignon, France; thomas.opitz@inra.fr*
[*] *Corresponding author*

Abstract. *Wildfires can cause important economic and ecological disasters. Their prevention begins with understanding the stochastic mechanisms governing the intensity of occurrences and the severity of fires. We focus on wildfires in the Mediterranean region Bouches-du-Rhone (South of France) observed since 1981, with burnt area larger than one hectare. Occurrences depend on the presence and concomitance of several factors: climatic (temperature, humidity, wind speed), environmental (vegetation types, urbanization, road network) and human activity. Whilst human activity is the main direct cause of wildfires, climatic and environmental conditions are a prior condition to their outbreak and propagation. Therefore, the structure of relative risk of wildfires is highly complex and shows strong variation over space and time and is driven by numerous covariates. Statistical challenges arise from the multi-scale spatio-temporal structure of data defined over various supports like fine grids for land use, coarse grids for fire position leading to positional uncertainty, and meteorological series observed at irregularly spaced measurement sites. The spatial heterogeneity of wildfires depends on the spatial distribution of current land use like vegetation type, urban zones or wetlands. We also show that changes in vegetation due to past fires affect the probability of wildfire occurrence during a regeneration period. Log-Gaussian Cox processes, along with the INLA method for inference and prediction, are particularly useful to model clustered events. Here, we show that they can also deal with more complex structures, allowing us to include the temporal inhibition at small spatial scales and thus providing more accurate predictions.*

Keywords. *Point processes; Spatial/spatio-temporal lattice data; Computational statistics; Environmental risk.*

---

1 Introduction

Since 1973, the French Government maintains the continuously updated Prométhée database[1] of wildfire occurrences in the Mediterranean region to develop statistical tools for spatial and temporal comparisons and a better knowledge of wildfire causes. Locations and dates of wildfires are recorded with associated characteristics such as the burnt area. We analyze and model wildfires in Bouches-du-Rhone (Southern France) since 1981 with burnt area larger than one hectare. Wildfire occurrence may depend on the presence and concomitance of several factors: climatic (temperature, humidity, wind speed), environmental

[1] `http://www.promethee.com/`

(vegetation types, urbanization, road network) and human activity. Whilst human activity is the main cause of wildfires, climatic and environmental conditions are a prior condition to their outbreak and their propagation.

The analysis of point patterns allows us to highlight the factors driving trends and interactions in the spatial distribution and the temporal structure of points and determine the observation scales of these relationships. The methods for analyzing and modeling point processes have been used in the context of wildfires [1, 2, 3, 4], but the spatial and temporal data have been treated separately or strongly aggregated (per year, per spatial area) which is unsatisfactory when one wants to understand and model the stochastic mechanisms of spatio-temporal interaction. However, some statistical methods exist for studying the spatio-temporal structures of such data [5, 6]; they yield a mechanistic or empirical modeling approach [7].

The spatio-temporal structure of the distribution of wildfires is very complex since, in practice, the dependence cannot be separated in space and time. The spatial heterogeneity of wildfires depends on the spatial distribution of current land use like vegetation, urban zones or wetlands [8, 9, 10, 2]. However, as it will be shown, it also depends on the past, because changes in vegetation due to fires affect the probability of wildfire occurrence during a regeneration period. In the literature, the hypothesis of separability in space and time is often assumed without any test because it allows decomposing the problem into two modeling steps, one in space and one in time, or to consider separable covariance matrices. This simplifying assumption led thus to progress on other scientific barriers.

We aim to detect and model multi-scale spatio-temporal structures in wildfire occurrences. Log-Gaussian Cox processes, along with the INLA method for inference and prediction, are particularly useful to model clustered events (see, *e.g.*, [10] and [2] in the context of wildfires). In this paper we show that they can also deal with more complex structures, allowing further temporal inhibition at small spatial scales and thus providing more accurate predictions.

2 Wildfire data and interaction of occurences over space and time

We consider a record of fire starting points for the years 1981 to 2015 for the Bouches-du-Rhone department in Southern France whose surface area amounts to around $5100km^2$. The spatial resolution is given by the DFCI coordinates spanning a grid in the Lambert 93 projection with quadratic grid cells covering approximately $4km^2$ each. The point coordinates of fires correspond to the center of the grid cell where the fire started. The value of burnt surface is also available for each event.

For weather data, we use freely available observation series from the Global Historical Climate Network (GHCN) hosted by the National Climatic Data Center[2] (NCDC). We here work with daily observation series of average temperatures, cumulated precipitation and maximum sustained wind speed for one measurement station (Marignane Airport, close to Marseille) in the Bouches-du-Rhone department.

We explore the influence of land use and climatic covariates like temperature and precipitation on the probability of event occurence.Also, we analyse interaction of wildfire occurences over space and time with the spatio-temporal inhomogeneous K-function defined in [6]. We observe spatio-temporal interaction, and particularly inhibition at small spatial distance certainly due to the absence of vegetation and other combustible material burnt after a wildfire.

[2] www.ncdc.noaa.gov/cdo-web/

3 Log-Gaussian Cox process models

We consider models with a stochastic intensity of log-Gaussian type and a space-year resolution for incorporating covariate information. Models of different complexity are considered for a Gaussian space-time effect $W(s,t)$, whose spatial component is always based on the flexible yet computationally convenient Matérn-like spatial Gauss–Markov random fields arising as approximate solutions to certain stochastic partial differential equations (see [11, 13] for details). The model is specified for years $t \in \{1981,\dots,2015\}$ in the following way:

$$\Lambda(s,t) = \exp\left(\beta_0 + \beta_{inhib} z_{inhib}(s,t) + \sum_{\text{land use}} \beta_{land,i} z_{land,i}(s) + \sum_{j=1}^{3} \beta_{clim,j} z_{clim,j}(t) + W(s,t)\right), \quad (1)$$

with covariates $z_{land,i}$ related to land use, $z_{clim,j}$ to climate and z_{inhib} to fires in the same DFCI cell during the 5 years preceding t.

We now explain the components in more detail. We have studied the following three structures for $W(s,t)$:

$$W(s,t) = W(s) \qquad \text{spatial marginal effect,} \quad (2)$$

$$W(s,t) = W(s) + W(t) \qquad \text{spatial and temporal marginal effects,} \quad (3)$$

$$W(s,t) = W_t(s) \qquad \text{spatial effects i.i.d. in time for } t = 1981,\dots,2015. \quad (4)$$

Models (2) and (3) can be considered either as a relatively simple log-Gaussian Cox processes, or as a Bayesian model for a Poisson process where the prior for the intensity is purely spatial in (2) and is space-time separable in (3). The marginal temporal effect $W(t)$ is here chosen as a first-order random walk with a sum-to-zero constraint. Model (4) incorporates higher stochasticity into the model through its replicated spatial effects; it is therefore capable to model clustering of events at the yearly level.

In models (2) and (3), a purely temporal effect is given through $\sum_{j=1}^{3} \beta_{clim,j} z_{clim,j}(t) + W(t)$, while $\sum_{\text{land use}} \beta_{land,i} z_{land,i}(s) + W(s)$ is a purely spatial effect. The inhibition effect $\beta_{inhib} z_{inhib}(s,t)$ is a novelty in our model compared to the existing literature and breaks the space-time separation of our model. It artificially integrates a repulsive pattern into the process, which remains well-defined since we can simulate the process iteratively for each time step by conditioning on the realization of the preceding time steps.

For estimating the posterior means of covariate coefficients and of the spatio-temporal effect $W(s,t)$, we use the framework of Integrated Nested Laplace Approximation [14], implemented in the `INLA` package of `R` [12]. Appropriate, only weakly informative priors were chosen for the estimated effects and hyperparameters like the effective range and the variance of the spatial Gaussian fields.

Estimation results and fire risk prediction will be shared during the conference.

References

[1] Genton, M., Butry, D., Gumpertz, M., and Prestemon, J. (2006). Spatio-temporal analysis of wildfire ignitions in the St Johns River water management district, Florida. *International Journal of Wildland Fire* **15**, 87–97.

[2] Serra, L., Saez, M., Mateu, J., Varga, D., Juan, P., Diaz-Avalos, C., and Rue, H. (2014). Spatio-temporal log-gaussian cox processes for modelling wildfire occurrence: the case of catalonia, 1994–2008. *Environmental and Ecological Statistics* **21**(3), 531–563.

[3] Turner, R. (2009). Point patterns of forest fire locations. *Environmental and Ecological Statistics*, (16), 197–223.

[4] Xu, H. and Schoenberg, F. (2011). Point process modelling of wildfire hazard in Los Angeles county, California. *The annals of Applied Statistics*, **The annals of Applied Statistics**, **5**, 684–704.

[5] Bonneu, F. (2007). Exploring and modeling fire department emergencies with a spatio-temporal marked point process. *Case Studies in Business, Industry and Government Statistics*, **1**, 139–152.

[6] Gabriel, E. and Diggle, P. (2009). Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica*, **63**, 43–51.

[7] Gabriel, E. (2016). Spatio-temporal point pattern analysis and modelling. In *Encyclopedia of GIS, 2nd Edition*.

[8] Juan, P., Mateu, J., and Saez, M. (2012). Pinpointing spatio-temporal interactions in wildfire patterns. *Stocastic Environmental Research and Risk Assesment*, **26**(8), 1131–1150.

[9] Moller, J. and Diaz-Avalos, C. (2010). Structured spatio-temporal shot-noise cox point process models, with a view to modelling forest fires. *Scandinavian Journal ofStatistics*, **37**(1), 2–25.

[10] Pereira, P., Turkman, K., Amaral-Turkman, M., Sa, A., and Pereira, J. (2013). Quantification of annual wildfire risk; a spatio-temporal point process approach. *Statistica*, **73**(1), 55–68.

[11] Lindgren, F., Rue, H., and Lindstrom, J. (2011). An explicit link between Gaussian fields and Gaussian Markov ranom fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(4), 423–498.

[12] Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, **63**(19).

[13] Simpson, D., Illian, J., Lindgren, F., Sorbye, S., and Rue, H. (2016). Going off grid: Computationally efficient inference for log-Gaussian Cox Processes. *Biometrika*. In press.

[14] Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, **71**(2):319–392.

A comparison of space-time estimation methods applied to air quality forecasting

M. Beauchamp[1,2*], L. Malherbe[2], C. de Fouquet [1], M. Valsania[3], F. Meleux[2] and A. Ung[2]

[1] *Mines ParisTech, Géosciences, Equipe géostatistique, 35 rue Saint Honoré, 77305 Fontainebleau, France; maxime.beauchamp@mines-paristech.fr, chantal.de_fouquet@mines-paristech.fr*
[2] *Institut National de l'Environnement Industriel et des Risques (INERIS), Direction des risques chroniques, Parc Technologique Alata, 60550 Verneuil-en-Halatte, France; laure.malherbe@ineris.fr, frederik.meleux@ineris.fr, anthony.ung@ineris.fr*
[3] *University of Turin, Via Verdi, 8 - 10124 Turin; marta.valsania@edu.unito.it*

Abstract. *The national PREV'AIR system (www2.prevair.org) delivers daily analyses and forecasts of different atmospheric pollutant concentrations over Europe and France. Forecast maps for the current and next two days (D+0, D+1, D+2) are computed by kriging statistical forecasts at the monitoring sites obtained by station specific multilinear regression models. Output data from the chemistry-transport model (CMT) CHIMERE are also used as an external drift in the kriging. In this study two kriging competitors are used: the usual space-time covariance-based kriging with external drift (KED) involving some appropriate neighbourhood to deal with reasonable CPU time and the SPDE-based kriging approach (SPDE). The performance is assessed using 2013 daily data and CMT simulations over France. It will be shown that both local fitting of the drift by (KED) and more global estimation made by (SPDE) can be a good alternative to the former (SA) framework.*

Keywords. *Prediction ; Generalized additive model ; Space-time kriging ; SPDE ; Air Quality*

Introduction

The prediction problem in Geosciences is often addressed through a data assimilation scheme to deal with the strong non-linearities of the underlying physical model. However, in Air Quality and because the forcing due to the emissions is often predominant over the quality of the estimation for the initial state, the so-called statistical adaptation that is a combination of local forecasts at the monitoring sites coupled with a spatial kriging with external drift has proved its good performing skill. In this work, usual covariance-based space-time kriging (KED) and SPDE-based kriging (SPDE) are confronted to (SA). The basic framework for each estimation method is first reminded. Then, some operational performance and cross-validation scores will be compared to the performance of the current statistical adaptation.

Space-time estimation methods

Let $Z(\mathbf{x}_\alpha, t_k),\ \alpha = 1, \cdots, N,\ k = 1, \cdots, M-1$ denote the space-time dataset of air quality daily concentrations observed at the monitoring sites $\mathbf{x}_\alpha$ between time t_1 and t_{M-1}, with possible missing values. Three estimations methods are used for the prediction problem, i.e. estimating the value at location $\mathbf{x}_0$ in the future $t_{M+l},\ l > 0$:

1) Space and Time independent estimation A generalized additive model is built for each monitoring sites $\mathbf{x}_\alpha$:

$$Z(\mathbf{x}_\alpha, t_k) = \beta_0 + \sum_{i=1,\cdots,p} f_i\big(\varphi_i(\mathbf{x}_\alpha, t_k)\big) + \varepsilon \tag{1}$$

where $\varphi_i(.,.),\ i = 1,\cdots,p$ are explanatory variables of the process $Z(.,.)$. The training dataset has to be long, several years if possible. A backfitting algorithm involving a 2 degree of freedom smoothing spline $\mathbf{S}_i$ to estimate f_i is used. The estimation at location $\mathbf{x}_0$ is given by a spatial kriging of the statistical forecasts obtained by these station specific gam models.

2) Covariance-based kriging approach $Z(\mathbf{x},t)$ is seen as a random function with deterministic part $\mu(\mathbf{x},t)$ and a residual $R(\mathbf{x},t)$:

$$Z(\mathbf{x},t) = \mu(\mathbf{x},t) + R(\mathbf{x},t) \tag{2}$$

A space-time kriging $Z(\mathbf{x},t) = \sum_{\alpha,\,k} \lambda_{\alpha,\,k} Z(\mathbf{x}_\alpha, t_k)$ is used for the estimation and the weights $\lambda_{\alpha,\,k}$ are solution of the linear system:

$$\begin{cases} \sum_{\alpha=1}^{n} \lambda_\alpha \gamma(\mathbf{x}_\alpha - \mathbf{x}_\beta, t_k - t_l) + \mu_0 + \sum_{i=1}^{p} \mu_i \varphi_i(\mathbf{x}_\beta, t_l) & = \gamma(\mathbf{x}_\beta - \mathbf{x}_0, t_k - t_0) \quad \forall \beta \\ \sum_{\alpha=1}^{n} \lambda_\alpha & = 1 \\ \sum_{\alpha=1}^{n} \lambda_\alpha \varphi_i(\mathbf{x}_\alpha, t_k) & = \varphi_i(\mathbf{x}_0, t_0) \quad \forall i \end{cases} \tag{3}$$

where $\gamma(.,.)$ denotes a space-time authorized variogram model, a Gneiting (Gneiting et al., 2007) or product-sum (De Iaco et al., 2001) model for instance. Because space-time datasets are large, an appropriate space-time neighbourhood has to be used, thus enabling a local fitting of the drift.

3) SPDE kriging approach Starting from the representation:

$$Z(\mathbf{x},t) = \sum_{k=1}^{n} \psi_l(\mathbf{x},t)\omega_k = \sum_{k=1}^{n} \psi_i^s(\mathbf{x})\psi_j^t(t)\omega_k$$

where the basis functions are seen as the product of purely spatial basis functions $\psi_i^s(\mathbf{s})$ and purely temporal basis functions $\psi_j^t(t)$, then the space-time stochastic PDE (Lindgren et al., 2011) defined by:

$$\frac{\partial}{\partial t}(\kappa(\mathbf{s})^2 - \Delta)^{\alpha/2}(\tau(\mathbf{s})x(\mathbf{s},t)) = \mathcal{W}(\mathbf{s},t), \quad (\mathbf{s},t) \in \mathcal{D} \times \mathbb{R}$$

generates a precision matrix $\mathbf{Q}$ for the Gaussian weights ω_k so that :

$$\mathbf{Q} = \mathbf{Q_T} \otimes \mathbf{Q_S}$$

$\mathbf{Q_s}$ and $\mathbf{Q_t}$ are respectively the precision matrices of the purely spatial model and the Markovian random walk, meaning that this model deals with separable covariances (Cameletti et al., 2012).

Results

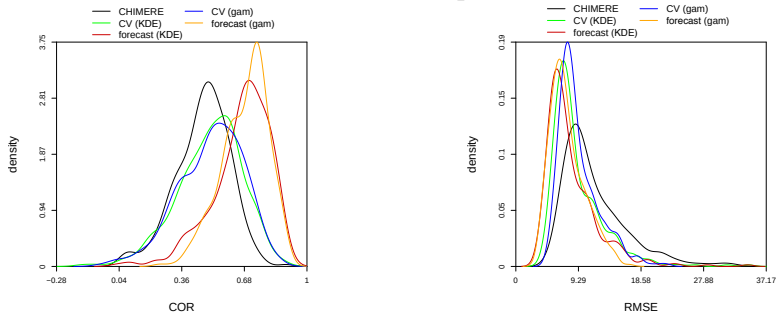
Some performance analyses are given for the daily prediction of PM_{10} and O_3 carried out in 2013. For kriging with external drift (KED), CHIMERE simulations are carried out over the FRA4k domain covering France and some parts of its neighbouring countries. For statistical adaptation (SA) and SPDE-based kriging (SPDE), some meteorological variables are also used, namely the daily temperature, average boundary layer height, and average specific humidity simulated by the meteorological model used by CHIMERE. (SA) also uses the daily concentration for $D-1$ as a predictor.

A look at cross-validation

Table 1 shows the performance, in terms of correlation, root mean square error (RMSE) and normalized mean bias (NMB) of the rough CHIMERE outputs, (SA) and (KED). Regarding these estimators, two ways of assessing the quality of the prediction is possible:
1) the cross-validation, i.e. to predict $Z(\mathbf{x}_\beta, t_0)$ the observational data for $D+0$ are not known. For (SA), the surrounding value of $Z(\mathbf{x}_\alpha, t_0)$, $\alpha \neq \beta$, are first estimated by their corresponding gam model before the spatial interpolation. For (KED), the kriging system is built upon the dataset $Z(\mathbf{x}_\alpha, t_k)\}$, $k \neq 0$, $\alpha \neq \beta$. In any case, for (SA) and (KED), the whole time series at $\mathbf{x}_\beta$ is removed so that this cross-validation will indicate the performance of the statistical estimation in area without any information in time and space.
2) the operational prediction score, i.e. the scores that would be obtained by comparing the predictions for $D+0$ made using all the data (including the values in the past at $\mathbf{x}_\beta$) to the observations collected the same day.

		daily data		
		Correlation	RMSE	NMB
	KED	0.48	9.48	5.2
PM_{10}	GAM	0.51	10.11	2.79
	KED	0.65	12.86	0.09
O_3	GAM	0.67	13.97	-1.22

Table 1: Cross-validation prediction score



(a) Correlation (PM_{10}) (b) RMSE (PM_{10})

Figure 1: Correlation and RMSE (PM_{10}) of CHIMERE outputs, the predictions made by (SA) and (KED), and the cross-validations made by (SA) and (KED)

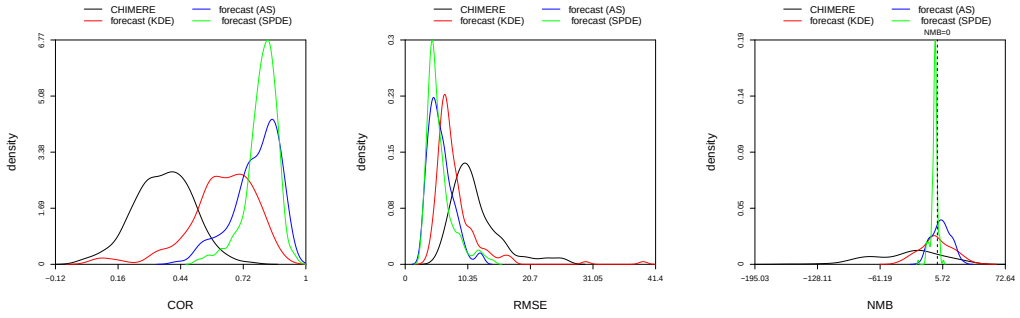
In areas well informed, with a high density of stations in the monitoring network, the prediction score indicates that using the gam model performs a bit better, which sounds normal because it benefits from the long training period of the model. At the same time, the cross-validation, which is a better indicator for assessing the quality of the final prediction map, shows that when the density of stations get lower, (KED) becomes more competitive and is better than (SA).

Operation prediction scores

Regarding (SPDE), only the operational performance score has been computed so far in the study. Though, it is still a good indicator for the performance of the prediction (see average scores in Table 2). To not increase the CPU time, about 2500 elements are used in the triangulation. Performance scores for PM_{10} on the whole second semester of 2013 reveals that SPDE-based kriging is now the most efficient. In addition, if by construction, (SA) and (KED) are unbiased, the variance of their errors is still quite large for space-time extrapolation. This variance is drastically reduced by the SPDE approach (see Figure 2).

		daily data		
		Correlation	RMSE	NMB
PM_{10}	KED	0.63	8.24	-1.27
	GAM	0.78	5.84	4.10
	SPDE	0.80	5.67	-3.13
O_3	KED	0.79	12.50	0.25
	GAM	0.88	8.99	0.85
	SPDE	0.86	9.46	-1.12

Table 2: Operational prediction scores



(a) Correlation (PM_{10}) (b) RMSE (PM_{10}) (c) NMB(PM_{10})

Figure 2: Correlation, RMSE and NMB (PM_{10}) of CHIMERE (AS), (KED) and (SPDE)

Conclusion

Both operational and cross-validation performance scores shows that space-time framework through kriging estimation is a solution to provide good predictions of the main air quality regulatory pollutants. The distinction between space and time in the statistical adaptation framework enables to introduce some sort of local non-stationarities that makes the prediction very good at the stations but its quality decreases quite fast when moving away from the available observations. Space-time kriging is more consistent regarding this point. Last, in the version of the SPDE approach, despite a modelling simpler for the covariance of the residuals than the one used in the usual covariance-based kriging, its performance is better. Probably because the drift is better estimated within this approach, which is finally more important than a better knowledge of the space-time structure of potentially larger residuals.

References

Cameletti, M.; Lindgren, F.; Simpson, D., and Rue, H. (2012). Spatio-temporal modeling of particulate matter concentration through the spde approach. *AStA Advances in Statistical Analysis*, 97(2):109–131. ISSN 1863-818X. doi: 10.1007/s10182-012-0196-3. URL http://dx.doi.org/10.1007/s10182-012-0196-3.

De Iaco, S.; Myers, D.E., and Posa, D. (2001). Space-time analysis using a general product-sum model. *Statistics & Probability Letters, v. 52, no. 1*, pages 21–28.

Gneiting, T.; Genton, M., and Guttorp, P. *Geostatistical Space-Time Models, Stationarity, Separability, and Full Symmetry*, pages 151–175. C&H/CRC Monographs on Statistics & Applied Probability. Chapman and Hall/CRC, (2007). ISBN 978-1-58488-593-1. doi: 10.1201/9781420011050.ch4. URL http://dx.doi.org/10.1201/9781420011050.ch4. 0.

Lindgren, F.; Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2011.00777.x. URL http://dx.doi.org/10.1111/j.1467-9868.2011.00777.x.

Evaluation of spatio-temporal Bayesian models for the spread of infectious diseases in oil palm

M. Denis[1,*], B. Cochard[2], I. Syahputra[3], H. de Franqueville[2] and S. Tisné[1]

[1] *UMR AGAP, CIRAD, Montpellier, France; marie.denis@cirad.fr, sebastien.tisne@cirad.fr*
[2] *PalmElit SAS, 34980 Montferrier sur Lez, France; benoit.cochard@palmelit.com*
[3] *P.T. Socfin Indonesia, Medan 20001, Indonesia; psbb_socfindo@yahoo.co.id*
[*] *Corresponding author*

Abstract. *In the field of epidemiology, studies are often focused on mapping diseases in relation to time and space. Hierarchical modeling is a common flexible and effective tool for modeling problems related to disease spread. In the context of oil palm plantations infected by the fungal pathogen* Ganoderma boninense*, we propose and compare two spatio-temporal hierarchical Bayesian models addressing the lack of information on propagation modes and transmission vectors. We investigate two alternative process models to study the unobserved mechanism driving the infection process. The models help gain insight into the spatio-temporal dynamic of the infection by identifying a genetic component in the disease spread and by highlighting a spatial component acting at the end of the experiment. In this challenging context, we propose models that provide assumptions on the unobserved mechanism driving the infection process while making short-term predictions using ready-to-use software.*

Keywords. *Bayesian modelling; Disease mapping; State-space models; Spatio-temporal models.*

---

1 Introduction

Basal stem rot caused by the fungal pathogen *Ganoderma boninense* is the major oil palm disease in Southeast Asia and is able to kill 80% of palm trees at the end of a planting cycle. While the struggle against *G. boninense* is a major concern in oil palm plantations, few studies have investigated the propagation mode or the genetic diversity and evolutionary history of the fungus. Two disease propagation modes have been documented to date: vegetative spread and basidiospore dissemination. But the importance of each mode and their relationships are not well known. Moreover while in most infectious diseases the transmission vectors are well established and helpful to understand the spread of the disease, *G. boninense* transmission vectors have not been well identified. Another factor involved in the infection spread concerns the genetic background of the host, which is crucial in the resistance to the infection. Most studies on *G. boninense* infection have shown a slow propagation with the apparition of the first symptoms after $4-5$ years. As a result, appropriate statistical models capable of addressing the lack of information on the disease propagation modes and transmission vectors are needed for analyzing spatially and/or temporally structured data. In our context of an oil palm plantation infected by *G. boninense* with little knowledge, two hierarchical Bayesian models were investigated to study unobserved mechanisms driving the infection process and to make short-term predictions using ready-to-use software. Both approaches are parameter-driven and based on the modeling of two stochastic processes

including spatial and temporal effects as well as interactions between space and time. The first approach directly integrates the status of the plot and neighborhood plots [9], while the second additively introduces spatial, temporal and spatio-temporal effects by using a tensor product B-splines [1]. The former may be compared to endemic-epidemic models, and the latter explores different types of interaction to capture the spatio-temporal dynamics of the infection. Binomial distribution will be considered instead of the commonly used Poisson distribution. Inference is achieved with the integrated nested Laplace approximation (INLA) approach as an efficient alternative to Markov chain Monte Carlo (MCMC) algorithms for inference [8]. In this talk, we present the first spatio-temporal analysis for the infection by *G. boninense*. The results provide a better understanding of the unobserved mechanism driving this infection

2 Methods

Let n_{stf} be the number of palm trees at risk in plot s $(s = 1,\ldots,S)$ at time t $(t = 1,\ldots,T)$ belonging to family f $(f = 1,\ldots,F)$. The proposed models can be formulated as hierarchical models:

1. **Data model:** $Y_{stf}|\pi_{stf} \sim \text{Binomial}(n_{stf},\pi_{stf})$, $s = 1,\ldots,S$, $t = 1,\ldots,T$ $f = 1,\ldots,F$ where y_{stf} is the number of newly infected palm trees for plot s belonging to family f at time t, and π_{stf} the associated probability.
2. **Process model:** $\log\left(\frac{\pi_{stf}}{1-\pi_{stf}}\right) = \eta_{stf} = \mu+\alpha_f+u_s+\theta_s+\gamma_t+\phi_t+\delta_{st}$, with μ denoting the intercept. α_f is a fixed effect corresponding to the effect being in family f relative to the referent family $f = 1$, with $\alpha_1 = 0$. u_s and θ_s correspond to unstructured and structured random spatial effects, respectively. Time effects defined by γ_t and ϕ_t can also be defined by different parameterizations. For a better understanding of the disease mapping, an interaction between space and time, $\delta = (\delta_{11},\ldots,\delta_{ST})$, is added in the equation of process model. Indeed only the main spatial and temporal effects are not enough for explaining differences in the temporal trend of infection risk for different geographical areas.
3. **Parameter models:** Parameter models refer to prior distributions for unknown parameters.

For both approaches the objectives are: (i) to model spatio-temporal data using relevant features for understanding the mechanism underlying the infection and, (ii) to be able to predict new infected trees. We use the deviance information criterion (DIC), and the Watanabe-Akaike information criterion (WAIC), an improvement on the DIC. For both, smaller values indicate a superior model. The probabilistic prediction performance can be evaluated using proper scoring rules [5]: the squared error score, the mean logarithmic of conditional predictive ordinate [7], and the logarithmic score.

2.1 Multivariate dynamic model

The first proposed model is based on a dynamic linear model. Dynamic models belong to the important class of state-space models and are mainly used in the context of an observable process depending on an unobserved state process. In the disease context, the mechanism driving the infection may be considered as an unobserved process. Based on [9], we use an autoregressive process for each plot, and an additional

component is integrated to model the spatial spread as a weighted sum of the past states in neighboring plots:

$$\eta_{stf} = \mu + \alpha_f + \beta t + \xi_{st}$$

$$\xi_{st} = \lambda \cdot \xi_{s,t-1} + \underbrace{\rho \cdot \sum_{s \neq s'} w_{s's} \cdot \xi_{s',t-1}}_{\text{spatial spread}} + \varepsilon_{st}$$

with μ denoting the intercept, α_f the effect of being in family f relative to the referent family $f = 1$, and β the regression coefficient associated with time t. An autoregressive process $\xi_s = (\xi_{s,1}, \ldots, \xi_{s,T})'$ governs the unobserved pattern of the spread in each plot s. The errors $\varepsilon = (\varepsilon_{11}, \ldots, \varepsilon_{ST})'$ are assumed to be independent and identically distributed normally distributed with variance σ^2_ε. The associated $w_{s's}$ weights are defined by 1 if two plots are neighbors, 0 otherwise.

2.2 B-spline model

Another approach considers an additive model for the process model. For minimizing the number of random effects resulting from the combination of spatial and temporal effects, the space-time interaction surface term is modeled with a tensor product B-spline [1]. The interaction term δ_{st} is defined as follows: $\delta_{st} = \sum_{k=1}^{K} b_{kt} B_{sk}$, $k = 1, \ldots, K$, with B_{sk} being the tensor product of two univariate B-spline functions of degree 3 (one for the “x", and one for the “y"), and K the number of basis functions. The spatio-temporal structure of the interaction term is specified through the prior distribution on the basis coefficients. Four interaction types are investigated [1, 3, 6]: 1) Type NoST: the product of prior distributions without any spatial and temporal structure for the basis coefficients,2) Type TnoS: the product of prior distributions with the temporal structure defined by a first-order random walk model and prior distributions without spatial structure, 3) Type SnoT: the product of an intrinsic conditional autoregressive (ICAR) prior [2] and prior distribution without any temporal structure, and 4) Type ST: the product of the RW1 prior and ICAR prior.

3 Application

Two modeling approaches are applied to analyze data on *G. boninense* infection in an oil-palm multi-parent population consisting of 14 full-sib families. This genetic trial was naturally infected and the infection status was recorded, most of time, biannually on 1,200 *Eg*9PP individuals over 25 years. The 14 full-sib families were represented with at least five replications, with each replication consisting of a plot with 15 full-sib individuals. In the following, data were aggregated by years and plots.
In terms of goodness-of-fit, B-spline models provide better fit to data compared to multivariate dynamic models. Using the tensor product of B-splines allows more complex modeling of interaction terms with weak sensitivity to the choice of basis number and locations compared to other tensor product choices. Thus this modeling helps gain insight into the spatio-temporal dynamics while yielding better results. The disadvantage of multivariate dynamic models in terms of goodness-of-fit may be due to an estimation of a common auto-regressive parameter for all plots. This limitation is due to the R-INLA package which does not allow consideration of the auto-regressive parameter as a random effect.
With regards to predictive performance, multivariate dynamic models performed better in most cases. Both modeling approaches captured the global infection dynamics. However multivariate dynamic models tended to smooth the estimations compared to B-spline models. In case of a peak in the percentage of infected palm trees, the prediction was more difficult for both approaches. As expected, the prediction

of the null percentage of infection was challenging for all models. The difficulty encountered in making short-term predictions by both approaches may be explained by the lack of knowledge on infection by *G. boninense* and by the real dataset that we considered.

4 Conclusion

In this study, two spatio-temporal hierarchical Bayesian approaches were investigated to model unobserved mechanisms driving the infection process due to *G. boninense* in oil palm plantations and to make short-term predictions with ready-to-use software. The main challenge was to address the lack of information on the disease propagation modes and transmission vectors by focusing on models with ready-to-use software. Two complementary models were proposed and provided first indications on the disease spread. Firstly, the comparison of different models revealed that the infection dynamics differed between families. Secondly, a spatial component involved at the end of the experiment was observed. Concerning the propagation modes, unfortunately, as the results were not able to distinguish the greatest propagation mode, further analyses with external factors are needed. The proposed models considered the neighborhood constant over time with the same effect for all neighbors. An interesting development would be to consider a dynamic neighbor selection, as proposed in [4], or/and to assume different weights $w_{ss'}$ over time and according to the infection levels. Although the *R-INLA* package permits high flexibility for modeling complex hierarchical models, these types of models cannot be fit. Other softwares such as JAGS or BUGS could be explored.

References

[1] Bauer, C., Wakefield, J., Rue, H., Self, S., Feng, Z., and Wang, Y. (2016). Bayesian penalized spline models for the analysis of spatio−temporal count data. *Statistics in medicine*, **35(11)**, 1848–1865.

[2] Besag, J., and Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 25–37.

[3] Blangiardo, M., and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wileyand Sons.

[4] Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A., and Schaap, M. (2016). Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *The Annals of Applied Statistics*, **10(3)**, 1286–1316.

[5] Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69(2)**, 243–268.

[6] Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in medicine*, **19**, 2555–2567.

[7] Pettit, L. I. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological)*, 175–184.

[8] Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, **71(2)**, 319–392.

[9] Schrodle, B., Held, L., and Rue, H. (2012). Assessing the impact of a movement network on the spatiotemporal spread of infectious diseases. *Biometrics*, **68(3)**, 736–744.

[10] Tisné, S., Pomiès, V., Riou, V., Syahputra, I., Cochard, B., and Denis, M. (2017). Identification of Ganoderma disease resistance loci using natural field infection of an oil palm multiparental population. *G3: Genes, Genomes, Genetics*, **7(6)**, 1683–1692.

Bayesian Measurement Error Correction in Structured Additive Distributional Regression with an Application to the Analysis of Sensor Data on Soil-Plant Variability

Giovanna Jona Lasinio[1,*], Alessio Pollice[2], Thomas Kneib[3],
Stefan Lang [4], Roberta Rossi[5], Mariana Amato[6]

[1] *Sapienza Università di Roma, Rome, ITALY; giovanna.jonalasinio@uniroma1.it*
[2] *Universitá degli Studi di Bari "Aldo Moro", Bari, ITALY ; alessio.pollice@uniba.it*
[3] *Universität Göttingen, Göttingen, GERMANY, tkneib@uni-goettingen.de*
[4] *Universität Innsbruck, Innsbruck, AUSTRIA, stefan.lang@uibk.ac.at*
[5] *CREA-ZOE , Bella (PZ), ITALY, roberta.rossi@crea.gov.it*
[6] *Università della Basilicata, Potenza, ITALY, mariana.amato@unibas.it*
[*] *Corresponding author*

Abstract. *The flexibility of the Bayesian approach to account for covariates with measurement error is combined with semiparametric regression models for a class of continuous, discrete and mixed univariate response distributions with potentially all parameters depending on a structured additive predictor. Markov chain Monte Carlo enables a modular and numerically efficient implementation of Bayesian measurement error correction based on the imputation of unobserved error-free covariate values. We allow for very general measurement errors, including correlated replicates with heterogeneous variances. The proposal is applied to the assessment of a soil-plant relationship crucial for implementing efficient agricultural management practices. Observations on multi-depth soil information and forage ground-cover for a seven hectares Alfalfa stand in South Italy were obtained using sensors with very refined spatial resolution. Estimating a functional relation between ground-cover and soil with these data involves addressing issues linked to the spatial and temporal misalignment and the large data size. We propose a preliminary spatial interpolation on a lattice covering the field and subsequent analysis by a structured additive distributional regression model accounting for measurement error in the soil covariate. Results are interpreted and commented in connection to possible Alfalfa management strategies.*

Keywords. *Bayesian modeling, computational statistics, semiparametric models.*

---

1 Introduction

Standard regression theory assumes that explanatory variables are deterministic or error-free, but this assumption is quite unrealistic for many biological processes and replicated observations of covariates are often obtained to quantify the variability induced by the presence of measurement error (ME). The most well known effect of measurement error is the bias towards zero induced by additive i.i.d. measurement error, but under more general measurement error specifications (as considered in this paper), different types of misspecification errors are to be expected [3, 6]. This is particularly true for semi-

parametric additive models, where the functional shape of the relation between responses and covariates is specified adaptively and therefore is also more prone to disturbances induced by ME. Recent papers advocate the hierarchical Bayesian modeling approach as a natural route for accommodating ME uncertainty in regression models. In this work we introduce a functional ME modeling approach allowing for replicated covariates with ME within a flexible class of regression models recently introduced, namely *structured additive distributional regression models* [4]. In this modeling framework, each parameter of a class of potentially complex response distributions is modeled by an additive composition of different types of covariate effects, e.g. non-linear effects of continuous covariates, random effects, spatial effects or interaction effects. We allow for quite general measurement error specifications including multiple replicates with heterogeneous dependence structure. From a computational point of view, based on the seminal work [2] for Gaussian scatterplot smoothing and [5] for general semiparametric exponential family and hazard regression models, we develop a flexible fully Bayesian ME correction procedure based on Markov chain Monte Carlo (MCMC) techniques to generate observations from the joint posterior distribution of structured additive distributional regression models. ME correction is obtained by the imputation of unobserved error-free covariate values in an additional sampling step. Our implementation is based on an efficient binning strategy that avoids recomputing the complete design matrix after imputing true covariate values and combines this with efficient storage and computation schemes for sparse matrices. The main motivation of our investigation comes from a case study on the use of proximal soil-crop sensor technologies to analyze the within-field spatio-temporal variation of soil-plant relationships in view of the implementation of efficient agricultural management practices. More precisely, we analyze the relationship between multi-depth soil information indirectly assessed through the use of high resolution geophysical soil proximal sensing technology and data of forage ground-cover variation measured by a multispectral radiometer within a seven hectares Alfalfa stand in South Italy.

2 Measurement Error Correction in Distributional Regression

The main motivation for our modeling proposal comes from the need to estimate the nonlinear dependence of ground-cover (NDVI) on soil information (Electro Resistivity, ER) by a smooth function, accounting for the heterogeneity in the position and scale of the response due to the sampling time, for the repeated measurements of the soil covariate and for the residual variation of unobserved spatial features.

2.1 Distributional Regression

Assume that independent observations (y_i, ν_i), $i = 1, \ldots, n$, are available on the response y_i and covariates ν_i and that the conditional distribution of the responses belongs to a K-parametric family of distributions such that $y_i|\nu_i \sim \mathcal{D}(\vartheta(\nu_i))$ and the K-dimensional parameter vector $\vartheta(\nu_i) = (\vartheta_1(\nu_i), \ldots, \vartheta_K(\nu_i))'$ is determined based on the covariate vector ν_i. More specifically, we assume that each parameter is supplemented with a regression specification $\vartheta_k(\nu_i) = h_k(\eta^{\vartheta_k}(\nu_i))$, where h_k is a response function that ensures restrictions on the parameter space and $\eta^{\vartheta_k}(\nu_i)$ is a regression predictor. In our analyses, we will consider one specific special case where $y_i \sim \text{Beta}(\mu(\nu_i), \sigma^2(\nu_i))$, i.e. responses are conditionally beta distributed with regression effects on location and scale. For both parameters $\mu(\nu_i)$ and $\sigma(\nu_i)^2$ of the beta distribution we employ a logit link, since they are restricted to the unit interval.

2.2 Structured Additive Predictor

For each of the predictors, we assume an additive decomposition as $\eta^{\vartheta_k}(\nu_i) = \beta_0^{\vartheta_k} + f_1^{\vartheta_k}(\nu_i) + \ldots + f_{J_k}^{\vartheta_k}(\nu_i)$, i.e. each predictor consists of a total of J_k potentially nonlinear effects $f_j^{\vartheta_k}(\nu_i)$, $j = 1,\ldots,J_k$, and an additional overall intercept $\beta_0^{\vartheta_k}$. The nonlinear effects $f_j^{\vartheta_k}(\nu_i)$ are a generic representation for a variety of different effect types (including nonlinear effects of continuous covariates, interaction surfaces, spatial effects, etc.). Any of these effects can be approximated in terms of a linear combination of basis functions as $f(\nu_i) = \sum_{l=1}^{L} \beta_l B_l(\nu_i) = b_i'\beta$, where we dropped both the function index j and the parameter index ϑ_k for simplicity, $B_l(\nu_i)$ denotes the different basis functions with basis coefficients β_l and $b_i = (B_1(\nu_i),\ldots,B_L(\nu_i))'$ and $\beta = (\beta_1,\ldots,\beta_l)'$ denote the corresponding vectors of basis function evaluations and basis coefficients, respectively. Since in many cases the number of basis functions will be large, we assign informative multivariate Gaussian priors $p(\beta|\theta) \propto \exp\left(-\frac{1}{2}\beta' K(\theta)\beta\right)$ to the basis coefficients to enforce certain properties such as smoothness or shrinkage. The specific properties are determined based on the prior precision matrix $K(\theta)$ which itself depends on further hyperparameters θ.

2.3 Measurement Error

In our application we are interested in estimating the nonlinear effect $f(x)$ in one of the predictors of a distributional regression model where instead of the continuous covariate x we observe M replicates $\tilde{x}_i^{(m)} = x_i + u_i^{(m)}, \quad m = 1,\ldots,M$, contaminated with measurement error $u_i^{(m)}$. For the measurement error, we consider a multivariate Gaussian model such that $u_i \sim \mathrm{N}_M(\mathbf{0}, \boldsymbol{\Sigma}_{u,i})$, where $u_i = (u_i^{(1)},\ldots,u_i^{(M)})'$ and $\boldsymbol{\Sigma}_{u,i}$ is a known, pre-specified unstructured covariance matrix. The basic idea in Bayesian measurement error correction is now to include the unknown, true covariate values x_i as additional unknowns to be imputed by MCMC simulations along with estimating the other parameters in the model. This requires that we assign a prior distribution to x_i as well and rely on the simplest version $x_i \sim \mathrm{N}(\mu_x, \tau_x^2)$, where we achieve flexibility by adding a further level in the prior hierarchy via $\mu_x \sim \mathrm{N}(0, \tau_\mu^2)$ and $\tau_x^2 \sim \mathrm{IG}(a_x, b_x)$. To obtain diffuse priors on these hyperparameters, we use $\tau_\mu^2 = 1000^2$ and $a_x = b_x = 0.001$ as default settings.

3 Case study

Given the aim of this work and the data size (ranging from 91438 to 222278 spatial points) , the spatial resolution was downscaled by interpolating samples to a 2574 cells square lattice overlaying the study area. Given the different number of sampled points corresponding to each sampling occasion (NDVI) and survey (ER), we used a proportional nearest neighbors neighborhood structure to compute the downscaled values. At each grid point we calculated the neighbors' means for both NDVI and ER, while neighbors' variances and covariances between depth layers were obtained for ER. Such a by-product of the downscaling of the original data is plugged into the model likelihood. For available NDVI recordings, we consider Beta distributional regression models and specify the two predictors as follows. For $s = 1,\ldots,2574$ grid points and $t = 1,\ldots,4$ time points, the structured additive predictor of the location parameter is determined as an additive combination of three linear and functional effects: a linear seasonal effect, a tensor product spatial effect and a nonlinear smooth effect of the continuous covariate ER. The linear predictor of the scale parameter is assumed to depend only on the effect of time, thus allowing heteroscedasticity of seasonal NDVI recordings. The Metropolis-Hastings algorithm, implemented using

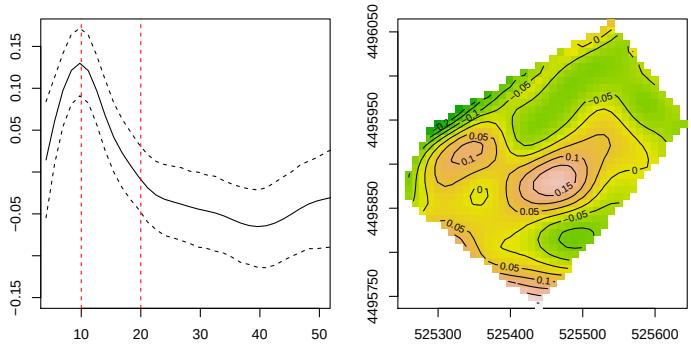


Figure 1: Smooth estimates of ER effects and residual spatial effects (both on the logit scale). Dotted vertical lines locate ER cut-offs corresponding to different monotonic soil-plant relationships.

`BayesX` [1], to sample the posteriors of the Beta models, required runs of 50000 iterations with 35000 burnin and thinning by 15. Convergence was reached and checked by visual inspection of the trace plots and standard diagnostic tools. Based on the resulting estimated smooth functions two ER cut-offs (at 10 and 20 *Ohm m*) are proposed that can be used to split the field in three areas characterized by a different monotonic soil-plant relationship: **Zone i: ER $<$ 10 *Ohm m***, where NDVI grows with ER and very low ER readings correspond to intermediate to high NDVI values; **Zone ii: 10 *Ohm m* $<$ ER $<$ 20 *Ohm m***, where ER is negatively related to NDVI and soil factors affecting ER act almost linearly and consistently on plant performance; **Zone iii: ER $>$ 20 *Ohm m***, where despite the large variation in ER there is a limited NDVI-soil responsiveness and NDVI is constantly low. Each zone conveys information on the shape and strength of the association between soil and crop variability, thus the proposed field zonation helps discerning areas where even a little change in soil properties can affect plant productivity (zone ii) from areas where soil environment is not practically alterable (zone iii) or in-season evaluations are possibly needed (zone i).

Acknowledgements The first two authors are partially supported by the MIUR-PRIN grant EphaStat (20154X8K23-SH3).

References

[1] Belitz, C., Brezger, A., Kneib, T., Lang, S. and Umlauf, N. (2015) , *BayesX: Software for Bayesian Inference in Structured Additive Regression Models*. Version 3.0.2.

[2] Berry, S. M., Carroll, R. J. and Ruppert, D. (2002) , 'Bayesian smoothing and regression splines for measurement error problems', *Journal of the American Statistical Association* **97**(457), 160–169.

[3] Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006) , *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*, Chapmann And Hall, CRC PRESS.

[4] Klein, N., Kneib, T., Klasen, S. and Lang, S. (2015) , 'Bayesian structured additive distributional regression for multivariate responses', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **64**(4), 569–591.

[5] Kneib, T., Brezger, A. and Crainiceanu, C. M. (2010) , Generalized semiparametric regression with covariates measured with error, *in* T. Kneib and G. Tutz, eds, 'Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir', Physica-Verlag HD, Heidelberg, pp. 133–154.

[6] Loken, E. and Gelman, A. (2017) , 'Measurement error and the replication crisis', *Science* **355**(6325), 584–585.

A Hierarchical Multivariate Spatio-Temporal Model for Clustered Climate data with Annual Cycles

Gianluca Mastrantonio[1,*], Giovanna Jona Lasinio[2], Alessio Pollice[3], Giulia Capotorti[4], Lorenzo Teodonio[5], Carlo Blasi[4]

[1] *Polytechnic of Turin; gianluca.mastrantonio@polito.it,*
[2] *DSS"Sapienza" University of Rome*
[3] *University of Bari "Aldo Moro"*
[4] *Department of environmental biology, "Sapienza" University of Rome*
[5] *ICRCPAL, Ministry of Cultural Heritage and Activities and Tourism, Roma*
[*] *Corresponding author*

Abstract. *We introduce a Bayesian multivariate hierarchical framework to estimate a space-time process model for a joint series of monthly extreme temperatures and amounts of rainfall. Data are available for 360 monitoring stations over 60 years, with missing data affecting almost all series. Model components account for spatio-temporal correlation and annual cycles, dependence on covariates and between responses. Spatio-temporal dependence is modeled by the nearest neighbor Gaussian process, response multivariate dependencies are represented by the linear model of coregionalization and effects of annual cycles are included by a circular representation of time. The proposed approach allows imputation of missing values and interpolation of climate surfaces at the national level. It also provides a characterization of the so called Italian ecoregions, namely broad and discrete ecologically homogeneous areas of similar potential as regards the climate, physiography, hydrography, vegetation and wildlife. To now, Italian ecoregions are hierarchically classified into 4 tiers that go from 2 Divisions to 35 Subsections and are defined by informed expert judgments. The current climatic characterization of Italian ecoregions is based on bioclimatic indices for the period 1955-2000.*

Keywords. *Bayesian modelling; Computational statistics; Geostatistics; Weather and Climate*

---

1 The problem

Climate elements and regimes, such as temperature, precipitation and their annual cycles, primarily affect the type and distribution of plants, animals, and soils as well as their combination in complex ecosystems [2]. The ecological classification of climate represents one of the basic steps for the definition and mapping of ecoregions, i.e. of broad ecosystems occurring in discrete geographical areas [1]. In keeping with these assumptions, a hierarchical classification of Italian ecoregions was recently obtained by informed expert judgments, including biogeography, physiography and climate among the main diagnostic features [3]. The Italian ecoregions are arranged into four hierarchically nested tiers, which consist of 2 Divisions, 7 Provinces, 11 Sections and 35 Subsections. The climatic features adopted for the diagnosis and description of the Italian ecoregions refer to bioclimatic indices that date back to the period 1955-2000. The primary focus of this work is the characterization of Italian ecoregions in terms of current

and past climatic conditions and involves summarizing climate variables at the ecoregion level, in order to evaluate climate impacts on ecosystems at the meso-scale and formulate reliable biodiversity conservation strategies. The secondary objective of our work is climate mapping. We address this issue by a fully model-based approach, relying on a stochastic model that accounts for some fundamental features of the multivariate spatio-temporal field that generates the data, i.e. correlation among climate variables and space-time variability. Estimation is embedded in the Bayesian hierarchical Gaussian modeling framework that allows control over various sources of uncertainty. While the richness and flexibility of spatio-temporal stochastic process models are indisputable, their computational feasibility and implementation pose some challenges for *large* datasets that are tackled using the nearest neighbor Gaussian process (NNGP) [5].

2 The data and the Model

Let $s \in \mathcal{S} \subset \mathbb{R}^d$, with $d = 2$, and $t \in \mathcal{T} \subset \mathbb{R}$ be spatial and temporal coordinates respectively, and let $Y_1^*(s,t)$, $Y_2^*(s,t)$ and $Y_3^*(s,t)$ represent the precipitation level, minimum and maximum temperature observed at (s,t). Then these variables have the following constraints: $Y_1^*(s,t) \geq 0$ and $Y_3^*(s,t) \geq Y_2^*(s,t)$. To simplify modeling and computations, we prefer to work with latent variables defined over the entire real line $\mathbb{R}$, embedding the above constraints in the variable definitions. Latent variables $Y_1(s,t)$, $Y_2(s,t)$ and $Y_3(s,t)$ are defined as follows:

$$\begin{cases} Y_1(s,t) = Y_1^*(s,t) & \text{if } Y_1^*(s,t) > 0, \\ Y_1(s,t) \leq 0 & \text{if } Y_1^*(s,t) = 0, \\ Y_2(s,t) = Y_2^*(s,t), & \\ Y_3(s,t) = Y_3^*(s,t) - Y_2^*(s,t) & \text{if } Y_3^*(s,t) - Y_2^*(s,t) > 0, \\ Y_3(s,t) \leq 0 & \text{if } Y_3^*(s,t) - Y_2^*(s,t) = 0. \end{cases}$$

Each latent response Y_i, $i = 1,2,3$ is described by a combination of fixed and random terms:

$$Y_i(s,t) = \mathbf{X}(s)\boldsymbol{\beta}_{z_k}(s) + \omega_i(s,t) + \lambda_i(s,t) + \varepsilon_i(s,t) \tag{1}$$

with $\varepsilon_i(s,t) \overset{iid}{\sim} N(0, \sigma^2_{\varepsilon,i})$. Here $\mathbf{X}(s) = (1, X(s))$ and $X(s)$ is the elevation of site s. The integer valued indicator $z_k(s) \subset \mathbb{Z}^+$ is the ecoregion label for the k^{th} ecoregion tier: with $k = 1$ we have one ecoregion covering the entire country, while $k = 5$ returns the finer classification with 35 ecoregions. The term $\lambda_i(s,t)$ describes the monthly effect of the annual cycle and we model it as

$$\lambda_i(s,t) \sim N\left(0, \sigma^2_{cy,i} \exp(-\phi_{cy,i} h_t^*)\right), \quad i = 1,2,3, \tag{2}$$

where $h_t^* = h_t \mod L$ is a circular distance with period $L = 1$ year. Finally, the random vector $\boldsymbol{\omega}(s,t) = (\omega_1(s,t), \omega_2(s,t), \omega_3(s,t))'$ is a multivariate spatio-temporal Gaussian process (GP) with dependent components, i.e $\boldsymbol{\omega}(s,t) = \mathbf{A}\mathbf{w}(s,t)$, where $\mathbf{w}(s,t) = (w_1(s,t), w_2(s,t), w_3(s,t))'$, $w_i(s,t) \perp w_j(s,t)$, for all (s,t)'s with $w_i \sim GP(\mathbf{0}, C(h_s, h_t; \boldsymbol{\theta}_i))$. For $C(h_s, h_t; \boldsymbol{\theta}_i)$ we choose the general non-separable space-time correlation structure [8]

$$C(h_s, h_t; \boldsymbol{\theta}_i) = \frac{1}{(\phi_{ti,i}|h_t|^{2\alpha_i} + 1)^{\tau}} \exp\left(-\frac{\phi_{sp,i}\|h_s\|^{2\gamma_i}}{(\phi_{ti,i}|h_t|^{2\alpha_i} + 1)^{\eta_i \gamma_i}} \right). \tag{3}$$

Remark that different specifications of matrix $\mathbf{A}$ can define different process structure [7]. In this work we use $\mathbf{A} = \boldsymbol{\Psi}\boldsymbol{\Gamma}\boldsymbol{\Psi}'$, where $\boldsymbol{\Gamma} = diag(\gamma_1, \gamma_2, \gamma_3)$ is the diagonal matrix of the square rooted eigenvalues of $\Sigma = \mathbf{A}\mathbf{A}'$ and $\boldsymbol{\Psi}$ is the orthogonal matrix of its eigenvectors; this choice guarantees that the process is invariant under reordering of the three observed variables.

		ϕ_{sp}	ϕ_t	ϕ_{cy}	η	σ^2_{cy}	σ^2_ε	σ^2_ω
Y_1	Est.	0.188	28.979	15.210	0.774	0.617	0.176	0.413
	(CI)	(0.184 0.192)	(28.871 29.072)	(14.972 15.394)	(0.774 0.775)	(0.612 0.623)	(0.175 0.178)	(0.409 0.416)
Y_2	Est.	0.138	9.628	10.176	0.943	6.968	0.008	0.050
	(CI)	(0.137 0.140)	(9.476 9.750)	(10.102 10.239)	(0.942 0.943)	(6.830 7.092)	(0.008 0.008)	(0.050 0.051)
Y_3	Est.	0.431	23.814	9.760	0.166	2.799	0.062	0.525
	(CI)	(0.429 0.432)	(23.576 23.995)	(9.690 9.835)	(0.165 0.168)	(2.647 2.919)	(0.061 0.062)	(0.519 0.532)

Table 1: Posterior estimates of the Gaussian process and annual cyclical component parameters.

	σ^2_ω	σ^2_{cy}	σ^2_ε
Y_1	0.342	0.512	0.146
Y_2	0.007	0.992	0.001
Y_3	0.155	0.827	0.018

Table 2: Proportions of the space-time, cyclical and residual components of the variance for each climate variable.

2.1 NNGP

To tackle the computational problems we use the NNGP. The basic idea is to write the joint density of the GP as the product of conditional densities. This multivariate density is approximated substituting each conditional set with smaller subsets containing at least m elements. More precisely

$$f(\boldsymbol{\omega}) = \prod_{n=1}^{N} f(\boldsymbol{\omega}_n|\boldsymbol{\omega}_{n-1},\ldots,\boldsymbol{\omega}_1) \approx \prod_{n=1}^{N} f(\boldsymbol{\omega}_n|\boldsymbol{\Omega}_n(m))$$

with $\boldsymbol{\omega}_0 = \emptyset$ and where $\boldsymbol{\Omega}_n(m) \subseteq (\boldsymbol{\omega}_{n-1},\ldots,\boldsymbol{\omega}_1)'$ is a subset that contains at most m elements of $(\boldsymbol{\omega}_{n-1},\ldots,\boldsymbol{\omega}_1)'$. As shown by [4], the quality of the approximation increases with m and the best results are achieved if we choose the m elements of $\boldsymbol{\Omega}_n$ that have the higher correlation with $\boldsymbol{\omega}_n$.

3 Model choice and posterior estimates

We estimated nine different models, varying the number of neighbors in the NNGP and the ecoregional hierarchical tier: $m = \{10,15,20\}$ and $k \in \{3,4,5\}$, respectively. Weakly informative priors were used throughout and the MCMC was implemented with 100000 iterations, a burn-in phase of 70000 and thinning by 12, keeping 2500 samples for posterior inferences. Posterior estimates were obtained in about three days and were implemented on the TeraStat cluster [6]. The choice among alternative specifications of the same model was performed using the DIC [9]. As expected, the largest number of neighbors always returns the smallest DIC value for a given k. In the following, we are going to report on parameter estimates and predictions obtained with the chosen model, that is the one with $m = 20$ and $k = 3$.
Posterior estimates of the GP and annual cyclical component parameters (see equations (2) and (3)) are reported in Table 1 with their 95% credible intervals (CI). In Table 2 we show the proportion of the variance due to the space-time, the cyclical and the residual component for each response variable, in order to appreciate the relevance of each of the three components in explaining the total variation.
Table 1 shows that the three climate variables have non-separable space-time dynamics, as CIs for the η parameter are never close to 0. Practical ranges and covariances of the three components in (1) provide useful information on the extent of the spatial, temporal and annual cyclical dependence. The spatial practical ranges of Y_1, Y_2 and Y_3 are respectively 15.95 *km*, 21.676 *km* and 6.967 *km*, while in terms of time dependence we have the following practical ranges: 37.78 days, 113.73 days and 45.98 days for Y_1, Y_2 and Y_3, respectively, and, finally, the annual cyclical effect ϕ_{CY} has similar behavior for the second and third variable, with practical ranges 71.99 days (Y_1), 107.60 days (Y_2) and 112.19 days (Y_3): annual

cycles are longer and almost seasonal (4 months long, as expected) for the minimum temperature and the temperature range, while a shorter cycle is estimated for the rain. Measures of the correlation between climate variables are also obtained and they are all far from zero.

4 Concluding remarks and future developments

The future will find us working on a more detailed bioclimatic characterization of the Italian ecoregions, obtaining parameter estimates for all available ecoregional tiers, including Divisions, Sections and Subsections. Further, as new ecoregional boundaries have recently been proposed mainly based on biogeographic and physiographic considerations (Blasi et al., unpublished data), the model could be applied to develop a climatic characterization of the new strata, comparing results to those reported in this paper.

Acknowledgements The first three authors are partially supported by the MIUR-PRIN grant EphaStat (20154X8K23-SH3).

References

[1] Bailey, R. G. (1983) , 'Delineation of ecosystem regions', *Environmental Management* **7**(4), 365–373.

[2] Bailey, R. G. (2004) , 'Identifying ecoregion boundaries', *Environmental Management* **34**(Suppl 1), S14–S26.

[3] Blasi, C., Capotorti, G., Copiz, R., Guida, D., Mollo, B., Smiraglia, D. and Zavattero, L. (2014) , 'Classification and mapping of the ecoregions of italy', *Plant biosystems - An International Journal Dealing with all Aspects of Plant Biology* **148**(6), 1255–1345.

[4] Datta, A., Banerjee, S., Finley, A. O. and Gelfand, A. E. (2016*a*) , 'Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets', *Journal of the American Statistical Association* **111**(514), 800–812.

[5] Datta, A., Banerjee, S., Finley, A. O. and Gelfand, A. E. (2016*b*) , 'On nearest-neighbor gaussian process models for massive spatial data', *Wiley Interdisciplinary Reviews: Computational Statistics* **8**(5), 162–171.

[6] Ferraro Petrillo, U. and Raimato, G. [2014) , 'Terastat computer cluster for high performance computing', "http://www.dss.uniroma1.it/en/node/6554" Department of Statistical Science Sapienza university of Rome.

[7] Gelfand, A. E., Schmidt, A. M., Banerjee, S. and Sirmans, C. F. (2004) , 'Nonstationary multivariate process modeling through spatially varying coregionalization', *TEST* **13**(2), 263–312.

[8] Gneiting, T. (2002) , 'Nonseparable, stationary covariance functions for space-time data', *Journal of the American Statistical Association* **97**(458), 590–600.

[9] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002) , 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4), 583–639.

Scale-Dependent Priors for Variance Parameters in Structured Additive Distributional Regression

N. Klein[1,*] and T. Kneib[2]

[1] *University of Cologne, Chair of Statistics, Universitaetsstr.22a, 50937 Cologne, n.klein@uni-koeln.de*
[2] *University of Goettingen, Chair of Statistics, Humboldtallee 3, 37073 Goettingen, tkneib@uni-goettingen.de*
[*] *Corresponding author*

Abstract. *The selection of appropriate hyperpriors for variance parameters is an important and sensible topic in all kinds of Bayesian regression models involving the specification of (conditionally) Gaussian prior structures where the variance parameters determine a data-driven, adaptive amount of prior variability or precision. We consider the special case of structured additive distributional regression where Gaussian priors are used to enforce specific properties such as smoothness or shrinkage on various effect types combined in predictors for multiple parameters related to the distribution of the response. Relying on a recently proposed class of penalised complexity priors motivated from a general set of construction principles, we derive a hyperprior structure where prior elicitation is facilitated by assumptions on the scaling of the different effect types. The posterior distribution is assessed with an adaptive Markov chain Monte Carlo scheme and conditions for its propriety are studied theoretically. We investigate the new type of scale-dependent priors in simulations and two challenging applications, in particular in comparison to the standard inverse gamma priors but also alternatives such as half-normal, half-Cauchy and proper uniform priors for standard deviations.*

Keywords. *Kullback Leibler divergence; Markov chain Monte Carlo simulations; Penalised complexity prior; Penalised splines; Propriety of the posterior.*

---

1 Introduction

Structured additive regression models are an important model class for regression modelling in various areas of applications. They combine the flexibility of generalised additive models with the inclusion of random effects, spatial components and further types of regression effects. While originally being restricted to responses from the exponential family, structured additive regression has recently been extended to a much broader class of response types known as distributional regression.
In these models, it is assumed that the (conditionally) independent response variables y_i, $i = 1,\dots,n$, given some covariate information $\boldsymbol{\nu}_i$ follow parametric distributions with density $p(y_i|\vartheta_{i1},\dots\vartheta_{iK})$ and distribution parameters ϑ_{ik}, $k = 1,\dots,K$. Each of the latter is linked to a structured additive predictor η_{ik} via a suitable one-to-one transformation h_k, i.e. $h_k(\eta_{ik}) = \vartheta_{ik}$. Dropping the parameter index k, the predictors are composed additively as

$$\eta_i = \beta_0 + \sum_{j=1}^{J} f_j(\boldsymbol{\nu}_i)$$

where, in turn, each function $f_j(\boldsymbol{\nu}_i)$ is represented by a linear combination of basis functions such that (suppressing i, j) $f(\boldsymbol{\nu}) = \sum_{d=1}^{D} \beta_d B_d(\boldsymbol{\nu})$. Here, $B_d(\boldsymbol{\nu}), d = 1, \ldots, D$, is a set of appropriate basis functions while $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_D)'$ is the vector of basis coefficients to be estimated.
To enforce specific properties such as smoothness we impose multivariate normal priors

$$p(\boldsymbol{\beta}|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta}\right)$$

with fixed positive (semi-)definite precision matrix $\mathbf{K}$ and variance parameter τ^2 that inherits the role of the smoothing parameter.
Suitable hyperpriors have then to be augmented to these variance components to complete the Bayesian model specification. While the inverse gamma prior $p(\tau^2) \sim \text{IG}(a, b)$ is a natural, conjugate prior, there has been considerable debate about the suitability of the inverse gamma distribution especially in the context of hierarchical random effects models. As a consequence, several alternatives such as half normal, half Cauchy or (proper) uniform priors for the standard deviation have been suggested as default priors in the literature [1]. Unfortunately, prior elicitation of the hyperparameters of these priors ensuring the propriety of the posterior and justification of the chosen distribution type with respect to axiomatic reasoning are often quite problematic in these cases.
Without relying on a specific modelling context, [2] develop a general approach for determining so-called penalised complexity priors reflecting that frequently hyperpriors are desired for parameters governing the deviation of a flexible model from a restrictive base model. We utilise this approach to develop scale-dependent priors for the variance parameters in distributional regression.

2 Scale-Dependent Hyperpriors

Omitting the distribution parameter index k as well as the effect index j we assume that $f(\boldsymbol{\nu})$ is one generic function of $\boldsymbol{\nu}$ included in a generic predictor $\boldsymbol{\eta}$.
The prior distribution $p(\tau^2|\theta)$ for the smoothing variances τ^2 constructed according to the principled definition of priors developed in [2] is a Weibull distribution with shape parameter $\alpha = 1/2$ and scale parameter θ, i.e. $p(\tau^2|\theta) = \frac{1}{2\theta}(\tau^2/\theta)^{-1/2}\exp\left(-(\tau^2/\theta)^{1/2}\right)$, as derived in the following:
1. *Occam's razor.* The prior invokes the principle of parsimony, i.e. a suitable base model for each effect is preferred as long as the data provide enough evidence for a more complex modelling.
2. *Measure of complexity.* The increased complexity between two models is measured by the unidirectional measure $d(p||p_b) = \sqrt{2\text{KLD}(p||p_b)}$ where $\text{KLD}(p||p_b)$ denotes the Kullback-Leibler divergence (KLD) between the base model represented by density p_b and the alternative represented by density p. Let $\text{N}(\mathbf{0}, \tau^2\mathbf{K}^-)$ denote the flexible model for a vector of regression coefficients and $\text{N}(\mathbf{0}, \tau_b^2\mathbf{K}^-)$ the base model with $\tau_b^2 \to 0$ and $\mathbf{K}^-$ the generalised inverse of $\mathbf{K}$. For $\tau^2 \gg \tau_b^2$ and $\tau_b^2 \to 0$ we then obtain $\text{KLD} \to \frac{\kappa}{2}\frac{\tau^2}{\tau_b^2}$ and hence a distance measure $d(\tau^2) = \sqrt{\frac{\kappa}{\tau_b^2}}(\tau^2)^{1/2}$.
3. *Constant rate penalisation.* This assumption implies an exponential prior on the distance scale $p_d(d) = \lambda\exp(-\lambda d)$ and finally delivers the prior on the original space as

$$p(\tau^2) = \lambda\exp(-\lambda d(\tau^2))\left|\frac{\partial d(\tau^2)}{\partial \tau^2}\right| = \frac{\lambda}{2}\sqrt{\frac{\kappa}{\tau_b^2}}(\tau^2)^{1/2}\exp\left(-\lambda\sqrt{\frac{\kappa}{\tau_b^2}}(\tau^2)^{1/2}\right).$$

Setting $\theta = (\lambda\sqrt{\frac{\kappa}{\tau_b 2}})^{-1/2}$ gives the prior above which we call scale-dependent hyperprior.

2.1 Choosing the Scale Parameter - User-Defined Scaling

The last principle controls the decay-rate $\exp(-\lambda)$ by imposing the condition $P(g(\tau^2) \leq c) = 1-\alpha$ for an interpretable transformation g of τ^2 and some user-defined values c and α. Compared to random effects model mostly considered in [2] we are interested in relating the scale parameter θ to the functions f rather than directly to the variances τ^2. This is achieved by specifying a certain interval the function f falls into with a high marginal probability

$$P(|f(x)| \leq c \; \forall x \in \mathcal{D}) \geq 1-\alpha$$

where $\alpha \in (0,1)$, $c > 0$ are chosen in advance and $\mathcal{D}$ is the domain of x. To solve the problem above, we use a finite subset $\mathcal{X}_P = \{x_1, \ldots, x_P\}$ of $\mathcal{D}$ together with the Bonferroni inequality to arrive at

$$P(|f(x_p)| \leq c \; \forall x \in \mathcal{X}_P) \geq 1 - \sum_{p=1}^{P} P\left(|f(x_p)| \geq c\right).$$

The marginal density of $f(x_p)$ can be obtained by integrating τ^2 out and the optimal scale parameter θ with respect to this criterion is obtained numerically.

3 Childhood Undernutrition in Zambia

As an illustration, we use data on 4421 children in Zambia. For each child i undernutrition is measured by a Z-score $zscore_i$ reflecting the nutritional status of child i with height h_i in the population of interest. The values m and s correspond to the mean height of children and their standard deviation in a suitable reference population of the same age group and gender.
We assume a location-scale model, that is, the Z-scores are conditionally normally distributed with

$$\begin{aligned} zscore_i &= \beta_0 + f_1(cage_i) + f_2(mage_i) + f_3(mbmi_i) + f_{spat}(district_i) + \varepsilon_i, \\ \varepsilon_i &\sim N(0, \sigma_i^2) \\ \log(\sigma_i^2) &= \tilde{\beta}_0 + \tilde{f}_1(cage_i) + \tilde{f}_2(mage_i) + \tilde{f}_3(mbmi_i) + \tilde{f}_{spat}(district_i). \end{aligned}$$

In the equations above, f_1 to f_3 ($\tilde{f}_1$ to $\tilde{f}_3$) are smooth functions of the continuous covariates *cage* (child's age), *mage* (mother's age at birth) and *mbmi* (mother's body mass index), while f_{spat} ($\tilde{f}_{spat}$) represents the spatial effect that was assigned a Markov random field prior, and β_0 ($\tilde{\beta}_0$) is the usual overall intercept. For all nonparametric effects, we use inverse gamma priors with default hyperparameters $a = b = 0.001$ and compare the results to scale-dependent priors.
Figure 1 is in accordance with the simulations and shows that the effects of *mage* and *mbmi* on the conditional mean are estimated to be closer to a linear effect with smaller credible intervals under the new scale-dependent prior. The spatial effect in Figure 2 is estimated similar with both hyperpriors which is reasonable in this case since the variable *district* has significant impact on both distribution parameters (based on a 95% credible interval).

References

[1] Gelman, A. (2006). Prior distributions for variance parameters in hierarchichal models *Bayesian Analysis*, **1**, 515–533.

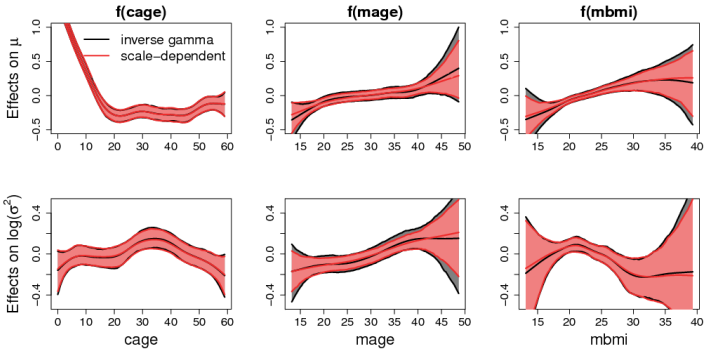


Figure 1: Comparison of estimated nonlinear effects of continuous covariates *cage*, *mage* and *mbmi*. Shown are posterior means and 95% credible intervals on $\mathrm{E}(zscore)$ (top) and on $\log(\sigma^2)$ (bottom).

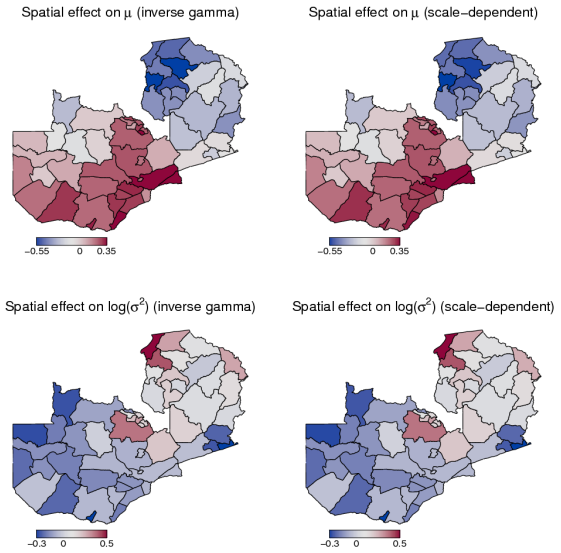


Figure 2: Comparison of estimated spatial effects for the inverse gamma prior (left) and scale-dependent prior (right). Shown are posterior means on $\mathrm{E}(zscore)$ (top) and on $\log(\sigma^2)$ (bottom).

[2] Simpson, D. P., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors *Statistical Science*, **1**, 1–28.

Spatio-Temporal Geostatistical Models associated to Evolution SPDEs

R. Carrizo Vergara [1,*], D. Allard[2] and N. Desassis[1]

[1] *Geostatistics, MINES ParisTech, 77350, Fontainebleau, France. ricardo.carrizo_vergara@mines-paristech.fr , nicolas.desassis@mines-paristech.fr*
[2] *BioSP, INRA PACA, Site Agroparc, 84914 Avignon Cedex 9, denis.allard@inra.fr*
[*] *Corresponding author*

Abstract. *We present a wide class of spatio-temporal geostatistical stationary models arising from spatio-temporal Stochastic Partial Differential Equations (SPDEs) interpreted in a distributional sense. We investigate the properties of models which are stationary solutions of first order and second order evolution SPDEs. When using a white noise as source term, remarkable properties about the spatio-temporal symmetry and the spatial behaviour have been obtained in the case of the first order models, and we show that these properties can be easily controlled by a suitable manipulation of the spatial operator of the SPDE. Second order models are always symmetric, and the spatial operator of the SPDE also determines its spatial structure. We detail important cases for these classes of models that have physical and statistical interest. In particular, we exemplify advection-diffusion equations, evolution of Matérn models, as well as the class of waving Matérn models which are stationary solutions of the homogeneous wave equation and usual Matérn models in space. We present simulations of these models, using algorithms based on PDE-solver, such as the finite difference method, the finite elements method, and Fourier methods.*

Keywords. *Spatio-temporal Covariance Functions; Space-Time Random Fields; Spectral Measure; Stochastic Partial Differential Equations; PDE-Solver Simulation*

---

1 Introduction

The Stochastic Partial Differential Equation (SPDE) approach has gained increasing interest in spatial statistics for its ability to model random fields with non-trivial properties [2] while offering very efficient estimation and simulation algorithms thanks to the use of PDE-solvers algorithms, as presented for example in [2] and [3].

In this work, we show that this approach can be made very general. It then allows to construct a large variety of models with special properties such as non-separability and spatio-temporal asymmetry. We present new classes of spatio-temporal geostatistical models which arise as stationary solutions for some evolution SPDEs. We describe their properties and their spatial behaviour. This class is quite general and can be related to a large variety of models, either already known or new. We provide examples related to the popular Matérn Model, and we show simulations based on PDE-solvers algorithms.

2 Evolution Models

We consider spatio-temporal stationary geostatistical models over $\mathbb{R}^d \times \mathbb{R}$ which are solutions of SPDEs of the form

$$\frac{\partial^n U}{\partial t^n} + \mathcal{L}_g U = X, \tag{1}$$

where $n = 1, 2$, X is a stationary (Generalized) Random Field called the source term, and $\mathcal{L}_g$ is a spatial-operator defined through the spatial symbol function g and the spatial Fourier transform and its inverse through $\mathcal{L}_g(\cdot) = \mathcal{F}_S^{-1}(g\mathcal{F}_S(\cdot))$. Under particular conditions on g and X, existence and uniqueness of stationary solutions of (1) can be guaranteed and characterized through their spectral measure (see more details in [1]).

As an example, the advection-diffusion equation (see [3]) is obtained for $n = 1$ by taking $g(\xi) = \kappa^2 + iv^T\xi + a|\xi|^2$ with $\kappa, a > 0$ and $v, \xi \in \mathbb{R}^d$. As a second example, the wave equation is obtained for $n = 2$ by taking $g(\xi) = c^2|\xi|^2$ with $c > 0$ and $\xi \in \mathbb{R}^d$.

2.1 First Order Evolution Models

First order evolution models correspond to $n = 1$ in (1). We focus on the case where X is a spatio-temporal white noise. Under suitable conditions on g (i.e., $|g|$ is inferiorly bounded by the reciprocal of a strictly positive polynomial), the spectral measure of the unique stationary solution is

$$d\mu_U(\xi, \omega) = \frac{1}{(2\pi)^{\frac{d+1}{2}}} \frac{d\xi d\omega}{(\omega + g_I(\xi))^2 + g_R^2(\xi)}, \quad (\xi, \omega) \in \mathbb{R}^d \times \mathbb{R} \tag{2}$$

and its covariance structure is described through the spatial Fourier transform $\mathcal{F}_S$:

$$\rho_U(h, u) = \mathcal{F}_S\left(\xi \mapsto \frac{1}{(2\pi)^{\frac{d}{2}}} \frac{e^{iug_I(\xi) - |u||g_R(\xi)|}}{2|g_R(\xi)|}\right)(h), \quad (h, u) \in \mathbb{R}^d \times \mathbb{R} \tag{3}$$

where g_R and g_I are respectively the real and imaginary part of g. The spatio-temporal asymmetry is controlled by g_I, with symmetry corresponding to $g_I = 0$. The spatial structure is determined by g_R, which can be seen by setting $u = 0$ in (3). The spatial behaviour of this class of models can thus be described through the SPDE over $\mathbb{R}^d$

$$\sqrt{2}\mathcal{L}_{\sqrt{|g_R|}} U_S = W_S, \tag{4}$$

where W_S is a spatial white noise. As a remarkable particular case, consider $g_R(\xi) = (\kappa^2 + |\xi|^2)^{\frac{\alpha}{2}}$ for $\kappa > 0$, $\alpha \in \mathbb{R}$. In this case, the spatial marginal random fields follow a Matérn model. We therefore refer to this class as *first order evolution of Matérn models*.

2.2 Second Order Evolution Models

Second order evolution models correspond to $n = 2$ in (1). When the source term is a spatio-temporal white noise, under suitable conditions on g, the spectral measure of a stationary solution is of the form

$$d\mu_U(\xi, \omega) = \frac{1}{(2\pi)^{\frac{d+1}{2}}} \frac{d\xi d\omega}{(\omega^2 - g_R(\xi))^2 + g_I^2(\xi)}, \quad (\xi, \omega) \in \mathbb{R}^d \times \mathbb{R}. \tag{5}$$

As this measure is temporally symmetric, we get a symmetric spatio-temporal model. The spatial structure of this model can be described through the SPDE over $\mathbb{R}^d$

$$\sqrt{2\sqrt{2}}\mathcal{L}_{|g|\sqrt{|g|-g_R}}U_S = W_S \tag{6}$$

When $g_R(\xi) = -(\kappa^2 + |\xi|^2)^{\frac{\alpha}{2}}$ and $g_I = 0$, we obtain models that are spatially Matérn. We therefore refer to this class as *second order evolution of Matérn models*.

As a second example of second order evolution model, one can show [1] that stationary solutions of the homogeneous wave equation

$$\frac{\partial^2 U}{\partial t^2} - c^2 \Delta U = 0. \tag{7}$$

have a covariance structure of the form

$$\rho_U(h,u) = \mathcal{F}_S\left(\xi \mapsto \cos(c|\xi||u|)d\mu_{U_S}(\xi)\right)(h), \quad (h,u) \in \mathbb{R}^d \times \mathbb{R} \tag{8}$$

where μ_{U_S} is the spectral measure of any arbitrarily chosen spatial stationary Random Field. For example, when $d\mu_{U_S}(\xi) = a(\kappa^2 + |\xi|^2)^{-\alpha}d\xi$, with $a, \kappa > 0$ and $\alpha \in \mathbb{R}$, we obtain a spatial Matérn covariance. The corresponding spatio-temporal model is referred to as the *waving Matérn model*.

3 Simulations

Simulations of these spatio-temporal models are performed by using PDE-solver algorithms, as presented in [2] and [3]. For the first order evolution models we apply the finite element method in space and the finite difference method in time to obtain simulations of some examples of first order evolution of Matérn models. For the waving Matérn model, we use spectral methods based on Fourier analysis, which are particularly adapted in the context of Stationary Random Fields.

References

[1] Carrizo Vergara R., Allard D., & Desassis N. (2018). A General Framework for Stationary Random Fields associated to SPDEs: Existence, Uniqueness and Examples. *Working Paper.*

[2] Lindgren F, & Rue H. (2011), An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B, 73*(4), 423-498, 2011.

[3] Sigrist, F., Künsch, H. R., & Stahel, W. A. (2015). Stochastic partial differential equation based modelling of large space-time data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77*(1), 3-33.

A family of random fields for modelling global data evolving in time: Regularity analysis and simulations

J. Clarke De la Cerda[1,*], A. Alegría[2] and E. Porcu[2]

[1] *CEREMADE, UMR CNRS 7534 Université Paris-Dauphine, PSL Research university; clarkemove@gmail.com*
[2] *Newcastle University, School of Mathematics and Statistics; alfredo.alegria.jimenez@gmail.com; georgepolya01@gmail.com*
[*] *Corresponding author*

Abstract. *In many geoscience applications, the phenomena of interest are observed on large portion of the Planet Earth. When such phenomena evolve over time, a valid model for the study of these observations is to consider them as partial realization of a spatio-temporal random field, where the spatial component is defined on a sphere, considering this last as a more realistic representation of the globe. We Introduce a family of Gaussian random fields (GRFs) constructed by spectral methods via a double Karhunen-Loève type representation. Based on recent result on the literature we claim that this family represent the class of Isotropic and stationary GRFs over $\mathbb{S}^d \times \mathbb{R}$ and study its regularity properties. In particular, we consider two alternative spectral decompositions for a GRF on $\mathbb{S}^d \times \mathbb{R}$. For each decomposition, we establish regularity properties through Sobolev and interpolation spaces. We then propose a simulation method and study its level of accuracy in the L^2 sense, which is fast and efficient.*
Then, we propose an extension of the previous results to GRF over the spherical shell (considering altitude), and to the longitudinally isotropic case. In both situations the random field is non-stationary in time.

Keywords. *Spatio-temporal Statistics; Gaussian random fields; Karhunen-Loève expansion; Longitudinally isotropic random fields; Spectral representation.*

---

1 Introduction and main result

Spatio-temporal variability is of major importance in many fields, in particular for anthropogenic and natural processes, such as earthquakes, geographic evolution of diseases, income distributions, mortality fields, atmospheric pollutant concentrations, hydrological basin characterization and precipitation fields, among others. For many natural phenomena involving, for instance, climate change and atmospheric variables, several branches of applied sciences have been increasingly interested in the analysis of data distributed over the whole sphere representing planet Earth and evolving through time. Besides, if we consider the variations in altitude, it is natural to consider the data as distributed over the spherical shell. Hence the need for random field models where the spatial location is continuously indexed through the sphere (or the spherical shell), and where time can be either continuous or discrete. It is common to consider the observations as a partial realization of a spatio-temporal random field which is usually considered to be Gaussian. Thus, the dependence structure in space-time is governed by the covariance

of a spatio-temporal Gaussian field.

Specifically, let d be a positive integer, and let $\mathbb{S}^d = \{\mathbf{x} \in \mathbb{R}^{d+1}, \|\mathbf{x}\| = 1\}$ be the d-dimensional unit sphere in the Euclidean space $\mathbb{R}^{d+1}$, where $\|\cdot\|$ denotes the Euclidean norm of $\mathbf{x} \in \mathbb{R}^{d+1}$. We denote $Z = \{Z(\mathbf{x},t),\ (\mathbf{x},t) \in \mathbb{S}^d \times \mathbb{R}\}$ a Gaussian field on $\mathbb{S}^d \times \mathbb{R}$.

The research to be presented extends part of the work of [5] to space-time. Such extension is non-trivial and depends on two alternative spectral decompositions of a Gaussian field on spheres cross time. In particular, we propose either Hermite or classical Karhunen-Loève expansions, show how regularity properties can evolve dynamically over time. Our main result proves that the smoothness of the covariance is related to the decay of the angular power spectrum. The crux of our arguments rely on recent advances on the characterization of covariance functions associated to Gaussian fields on spheres cross time (see [1]) and the ideas introduced in [4].

Later, we introduce a simulation method for $d = 2$ which is computationally fast while keeping a reasonable level of accuracy, resulting in a notable step forward. The method is based on a suitable truncation of the proposed double spectral decompositions. The main result of this section is to establish the accuracy of the method in the L^2 sense, then we illustrate how the model keeps a reasonable level of precision while being considerably fast, even when the number of spatio-temporal locations is very high.

Let Z be a random field on $\mathbb{S}^d \times \mathbb{R}$ defined, in the mean square sense, as

$$Z(\mathbf{x},t) = \sum_{j=0}^{\infty} \sum_{m=0}^{\dim(\mathcal{H}_j^d)} X_{j,m,d}(t)\mathcal{Y}_{j,m,d}(\mathbf{x}). \tag{1}$$

Here, for each j,m and d, $\mathcal{Y}_{j,m,d}$ are the Spherical Harmonics basis functions, $\mathcal{H}_j^d$ denote the linear space of spherical harmonics of degree j over $\mathbb{S}^d$, and $\{X_{j,m,d}(t), t \in \mathbb{R}\}$ is a complex-valued zero-mean stationary Gaussian process such that

$$\begin{aligned}\operatorname{cov}\{X_{j,m,d}(t), X_{j',m',d}(s)\} &:= \mathbb{E}\{X_{j,m,d}(t)\ \overline{X_{j',m',d}(s)}\} \\ &= \varphi_{j,d}(t-s)\delta_{j,j'}\delta_{m,m'},\end{aligned}$$

where $\{\varphi_{j,d}\}_{j\in\mathbb{N}}$ represents the Schoenberg's functions associated to the covariance kernel mapping ψ of the process Z,

$$\psi(x,t-s) = \operatorname{cov}\{Z(\mathbf{x},t), Z(\mathbf{y},s)\} = \sum_{j=0}^{\infty} \varphi_{j,d}(t-s)c_j(d,x)\dim(\mathcal{H}_j^d), \quad x \in [-1,1],\ \ (t-s) \in \mathbb{R}. \tag{2}$$

Here, $c_j(d,x)$ are the standardized Gegenbauer polynomials, and the series is uniformly convergent.

By applying a second Karhunen-Loève decomposition to the process Z, an alternative way to write (1) is:

$$Z(\mathbf{x},t) = \sum_{j=0}^{\infty} \sum_{m=1}^{\dim(\mathcal{H}_j^d)} \sum_{k=0}^{\infty} a_{j,k,m,d}\,\zeta_k(t)\mathcal{Y}_{j,m,d}(\mathbf{x}), \quad (\mathbf{x},t) \in \mathbb{S}^d \times \mathbb{R}. \tag{3}$$

Expressions (1) or (3) represent a way to construct isotropic stationary GRFs on the sphere cross time, and suggest a spectral simulation method, which consists in truncating the double series.

1.1 Simulations

For $d = 2$, simple examples can be generated from the following space-time angular power spectrum

$$a_{j,k} = \frac{1}{1+(1+j)^{\nu_1}(1+k)^{\nu_2}}, \tag{4}$$

with $\nu_i > 2$, for $i = 1,2$. We illustrate space-time realizations on $\mathbb{S}^2 \times \{1,2\}$, over 24000 spatial locations, with coefficients (4), in two cases: (a) $\nu_1 = 3$ and $\nu_2 = 5$, and (b) $\nu_1 = \nu_2 = 5$.

Figures 1 and 2 show the corresponding realizations for cases (a) and (b), respectively. For each case, we truncate the series (3) for $d = 2$ using $K = J = 50$. Note that the parameter ν_1 is the responsible of the spatial scale and smoothness of the realization. In [5], some realizations are illustrated using a similar spectrum, in a merely spatial context.

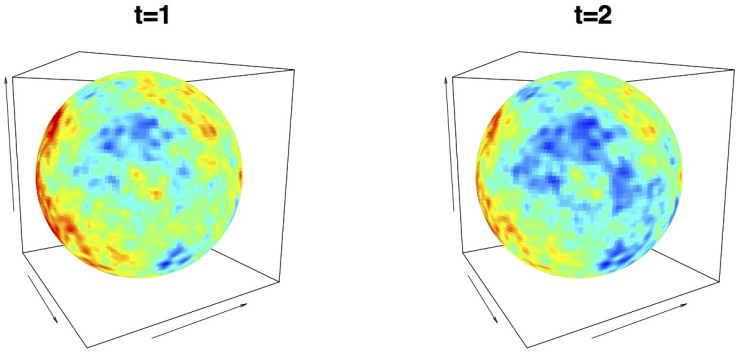


Figure 1: Space-time realization on $\mathbb{S}^2 \times \{1,2\}$, with spectrum (4), with $\nu_1 = 3$ and $\nu_2 = 5$.

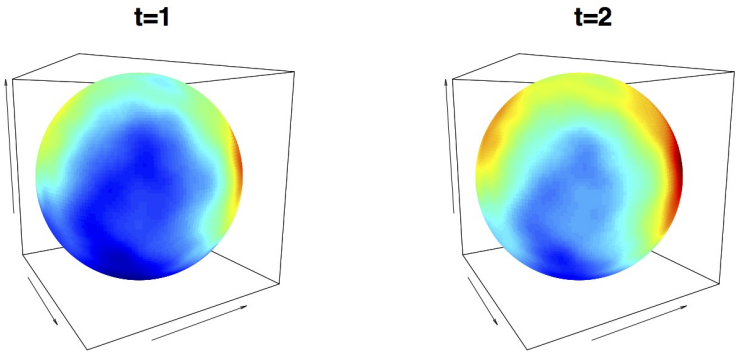


Figure 2: Space-time realization on $\mathbb{S}^2 \times \{1,2\}$, with spectrum (4), with $\nu_1 = \nu_2 = 5$.

1.2 Extension of the results

As mentioned in the introduction, we will present partial extensions of the results introduced in [2], to two different scenarios, both situations being non-stationary in the time variable:

i.- GRFs on the Spherical Shell cross time,

ii.- Longitudinally isotropic GRFs on the sphere cross time.

The first case is motivated by a real data-set of temperature and humidity, measured over the whole earth from the sea level until 30000m of altitude. We have decided to consider a more realistic approach for this kind of data, under the paradigm of spatio-temporal statistics modelling. This time, our objective is to obtain similar results as those presented in [2], but for random fields with spatial locations over the spherical shell $\mathbb{S}^d_{r_1,r_2} := \{\mathbf{x} \in \mathbb{R}^{d+1} : 0 \leq r_1 \leq \|\mathbf{x}\|_2 \leq r_2 < +\infty\}$. This will allow to differentiate the variations of the phenomena in longitude, latitude and altitude. Besides, thanks to recent results on positive definite functions (see [3]), we only need to assume a condition of isotropy in latitude/longitude, thus going beyond stationarity in time or altitude.

For the second case, the inspiration comes from several phenomena evolving over a large part of the Earth, or all over the Earth, that have a strong correlation along the longitudes, but very weak throughout the latitudes, in particular the cases of aerosol dynamics. This prompted us to seek an alternative to the axially symmetric random fields as the theory for the anisotropic processes on the sphere.

References

[1] Berg, C. and Porcu, E. (2016), From Schoenberg coefficients to Schoenberg functions. *Constr. Approx.*, 1–25.

[2] Clarke De la Cerda, J. Alegría, A. and Porcu, E. (2018), Regularity properties and simulations of Gaussian random fields on the sphere cross time, 2018. *Electron. J. Stat.*, 12 (1), 399–426.

[3] Guella J. and Menegatto V. (2018), From Schoenberg coefficients to Schoenberg functions: a unifying framework. *Submitted*. http://conteudo.icmc.usp.br/pessoas/menegatt/prepa.html

[4] Jones, R. H. (1963), Stochastic processes on a sphere. *Ann. Math. Statist.*, **34**, 213–218.

[5] Lang, A. and Schwab, C. (2015), Isotropic Gaussian random fields on the sphere: regularity, fast simulation and stochastic partial differential equations. *Ann. Appl. Probab.*, **25**, 3047–3094.

Constructing a Spatial Concordance Correlation Coefficient

Ronny Vallejos[1,*], Javier Pérez[2], Aaron Ellison[3] and Andrew D. Richardson[4]

[1,2] *Departamento de Matemática, Universidad Técnica Federico Santa María, Avenida España 1680, Valparaíso, Chile; ronny.vallejos@usm.cl, javier.perez@alumnos.usm.cl,*
[3] *Harvard Forest, Harvard University, Petersham, Massachusetts, USA; aellison@fas.harvard.edu*
[4] *School of Informatics, Computing and Cyber Systems, Northern Arizona University, USA and Center for Ecosystem Science and Society, Northern Arizona University, USA; Andrew.Richardson@nau.edu*
[*] *Corresponding author*

Abstract. *In this work we define a spatial concordance coefficient for second-order stationary processes. This problem has been widely addressed in a non-spatial context, but here we consider a coefficient that for a fixed spatial lag allows one to compare two spatial sequences along a 45° line. The proposed coefficient is explored for the bivariate Matérn and Wendland covariance functions. The asymptotic normality of a sample version of the spatial concordance coefficient for an increasing domain sampling framework is established for the Wendland covariance function. Monte Carlo simulations are used to gain additional insights into the asymptotic properties for finite sample sizes. The results will be illustrated by real data examples to see how our method works in practice.*

Keywords. *Concordance; Correlation; Spatial correlation function; Lin's coefficient; Bivariate Wendland covariance function.*

---

1 A Concordance Correlation Coefficient

In recent decades, concordance correlation coefficients have been developed in a variety of different contexts. For instance, in assay or instrument validation processes, the reproducibility of the measurements from trial to trial is of interest. Also, when a new instrument is developed, it is relevant to evaluate whether its performance is concordant with other, existing ones. In the literature, this concordance has been tackled from different perspectives [1]. One way to approach this problem for continuous measurements is constructing a scaled summary index that can take on values between -1 and 1. Using this approach Lin [6] suggested a concordance correlation coefficient that evaluates the agreement between two continuous variables by measuring the variation from a 45° line through the origin.

More precisely, assume that X and Y are two continuous random variables such that the joint distribution of X and Y has finite second moments with means μ_X and μ_Y, variances σ_X^2 and σ_Y^2, and covariance σ_{YX}. The mean squared deviation of $D = Y - X$ is

$$\text{MSD} = \varepsilon^2 = \mathbb{E}[D^2] = \mathbb{E}[(Y-X)^2].$$

It is straightforward to see that $\varepsilon^2 = (\mu_X - \mu_Y)^2 + \sigma_Y^2 + \sigma_X^2 - 2\sigma_{YX}$ and the sample counterpart satisfies $e^2 = (\overline{y} - \overline{x})^2 + s_Y^2 + s_X^2 - 2s_{XY}$. Under the above hypothesis, Lin [6] proposed a concordance correlation

coefficient defined as

$$\rho_c = 1 - \frac{\varepsilon^2}{\varepsilon^2|\rho=0} = \frac{2\sigma_{YX}}{\sigma_Y^2+\sigma_X^2+(\mu_Y^2-\mu_X^2)^2}. \quad (1)$$

This coefficient satisfies the following properties:

1. $\rho_c = \alpha\cdot\rho$, where $\alpha = \frac{2}{w+1/w+v^2}$ and $w = \frac{\sigma_Y}{\sigma_X}$.
2. $|\rho_c| \leq 1$.
3. $\rho_c = 0$ if and only if $\rho = 0$.
4. $\rho_c = \rho$ if and only if $\sigma_Y = \sigma_X$ and $\mu_Y = \mu_X$.

The sample estimate of ρ_c is given as

$$\widehat{\rho}_c = \frac{2s_{YX}}{s_Y^2+s_X^2+(\bar{y}-\bar{x})^2}.$$

The inference for this coefficient was addressed via Fisher's transformation. Lin [6] proved that

$$Z = \frac{1}{2}\left(\frac{1+\widehat{\rho}_c}{1-\widehat{\rho}_c}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(\psi,\sigma_Z^2), \text{ as } n\to\infty,$$

where $\psi = \tanh^{-1}(\rho_c) = \frac{1}{2}\left(\frac{1+\rho_c}{1-\rho_c}\right)$, $\sigma_Z^2 = \frac{1}{n-2}\left[\frac{(1-\rho^2)\rho_c^2}{(1-\rho_c^2)\rho^2} + \frac{2v^2(1-\rho_c)\rho_c^3}{(1-\rho_c^2)^2\rho} + \frac{v^4\rho_c^4}{2(1-\rho_c^2)^2\rho^2}\right]$, and $v^2 = \frac{(\mu_Y-\mu_X)^2}{\sigma_Y\sigma_X}$. As a consequence of the asymptotic normality of the sample concordance index, an approximate hypothesis testing problem of the form $H_0 : \rho_c = \rho_0$ versus $H_1 : \rho_c \neq \rho_0$ for a fixed ρ_0 can be constructed.

Applications and extensions of Lin's coefficient can be found in [5], [7], and [8], among others.

2 A Spatial Concordance Coefficient and its Properties

Here, we extend Lin's coefficient for bivariate second-order spatial processes for a fixed lag in space.

Definition 1. Let $(X(\boldsymbol{s}),Y(\boldsymbol{s}))^\top$ be a bivariate second-order stationary random field with $\boldsymbol{s}\in\mathbb{R}^2$, mean $(\mu_1,\mu_2)^\top$, and covariance function

$$C(\boldsymbol{h}) = \begin{pmatrix} C_X(\boldsymbol{h}) & C_{XY}(\boldsymbol{h}) \\ C_{YX}(\boldsymbol{h}) & C_Y(\boldsymbol{h}) \end{pmatrix}.$$

Then the spatial concordance coefficient is defined as

$$\rho^c(\boldsymbol{h}) = \frac{\mathbb{E}[(Y(\boldsymbol{s}+\boldsymbol{h})-X(\boldsymbol{s}))^2]}{\mathbb{E}[(Y(\boldsymbol{s}+\boldsymbol{h})-X(\boldsymbol{s}))^2|C_{XY}(\boldsymbol{0})=0]} = \frac{2C_{YX}(\boldsymbol{h})}{C_X(\boldsymbol{0})+C_Y(\boldsymbol{0})+(\mu_1-\mu_2)^2}. \quad (2)$$

Some straightforward features of this coefficient are the following:

1. $\rho^c(\boldsymbol{h}) = \eta\cdot\rho_{YX}(\boldsymbol{h})$, where $\eta = \frac{2\sqrt{C_X(\boldsymbol{0})C_Y(\boldsymbol{0})}}{C_X(\boldsymbol{0})+C_Y(\boldsymbol{0})+(\mu_1-\mu_2)^2}$.
2. $|\rho^c(\boldsymbol{h})| \leq 1$.
3. $\rho^c(\boldsymbol{h}) = 0$ iff $\rho_{YX}(\boldsymbol{h}) = 0$.
4. $\rho^c(\boldsymbol{h}) = \rho_{YX}(\boldsymbol{h})$ iff $\mu_1 = \mu_2$ and $C_X(\boldsymbol{0}) = C_Y(\boldsymbol{0})$.

5. For a bivarite Matérn covariance function defined as $C_X(\boldsymbol{h}) = \sigma_1^2 M(\boldsymbol{h}, \nu_1, a_1)$, $C_Y(\boldsymbol{h}) = \sigma_2^2 M(\boldsymbol{h}, \nu_2, a_2)$, $\mu_1 = \mu_2$, $C_{YX}(\boldsymbol{h}, \nu_{12}, a_{12}) = \rho_{12}\sigma_1\sigma_2 M(\boldsymbol{h}, \nu_{12}, a_{12})$, where $M(\boldsymbol{h}, \nu, a) = (a||\boldsymbol{h}||)^{\nu} K_{\nu}(a||\boldsymbol{h}||)$, and $K_{\nu}(\cdot)$ is a modified Bessel function of the second type and $\rho_{12} = \text{cor}[X(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)]$ we have that

$$\rho^c(\boldsymbol{h}) = \frac{2\sigma_1\sigma_2 M(\boldsymbol{h}, \nu_{12}, a_{12})}{\sigma_1^2 + \sigma_2^2} = \eta \cdot \rho_{12},$$

where $\eta = \frac{2\sigma_1\sigma_2 M(\boldsymbol{h}, \nu_{12}, a_{12})}{\sigma_1^2 + \sigma_2^2}$.

6. For a bivariate Wendland-Gneiting covariance function [3] of the form

$$\boldsymbol{C}(\boldsymbol{h}) = [\rho_{ij}\sigma_{ii}\sigma_{jj} R_{ij}(\boldsymbol{h})]_{i,j=1}^2,$$

where $R(\boldsymbol{h}, \psi_{12}) = c_{ij} b_{ij}^{\nu+2k+1} B(\nu + 2k + 1, \gamma_{ij} + 1) \tilde{\psi}_{\nu+\gamma_{ij}+1,k}\left(\frac{\|\boldsymbol{h}\|}{b_{ij}}\right)$, $B(\cdot,\cdot)$ is the beta function, and $\tilde{\psi}_{\nu,k}$ is defined in [4], the spatial concordance coefficient is

$$\rho^c(\boldsymbol{h}) = \frac{2\rho_{12}\sigma_1\sigma_2 R(\boldsymbol{h}, \psi_{12})}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}, \quad \boldsymbol{h} \in \mathbb{R}^2, \tag{3}$$

In particular, considering $R_{ij}(\boldsymbol{h}) = p_k(\|\boldsymbol{h}\|)(1 - \|\boldsymbol{h}\|/b_{ij})_+^l$, where $k = 2$, $l = v + 1$, and $b_{ij} > 0$,

$$\rho^c(\boldsymbol{h}) = \frac{2\rho_{12}\sigma_1\sigma_2 (1 + l\|\boldsymbol{h}\|/b_{12})(1 - \|\boldsymbol{h}\|b_{12})_+^l}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)}.$$

3 Inference

In the previous section we proved that for several covariance structures, the spatial concordance coefficient defined in (2) can be written as a product of the correlation coefficient and a constant. Thus, we can consider a plug-in estimator for these two quantities.

Let $(Z_1(\boldsymbol{s}), Z_2(\boldsymbol{s}))^{\top}$ $s \in D$ be a Gaussian process with mean $\boldsymbol{\mu} = (\mu_1, \mu_2)^{\top}$ and covariance function $\boldsymbol{C}(\boldsymbol{h})$, $\boldsymbol{s}, \boldsymbol{h} \in \mathbb{R}^2$. Then a sample estimate of the concordance index (2) is

$$\widehat{\rho}_c(\boldsymbol{h}) = \widehat{\rho}_{12}(\boldsymbol{h})\widehat{C}_{ab}, \tag{4}$$

where $\widehat{C}_{ab} = ((\widehat{a} + 1/\widehat{a} + \widehat{b}^2)/2)^{-1}$, $\widehat{a} = \left(\frac{\widehat{C}_{11}(\boldsymbol{0})}{\widehat{C}_{22}(\boldsymbol{0})}\right)^{1/2}$, $\widehat{b} = \frac{\widehat{\mu_1} - \widehat{\mu_2}}{(\widehat{C}_{11}(\boldsymbol{0})\widehat{C}_{22}(\boldsymbol{0}))^{1/4}}$ and $\widehat{\mu}_1, \widehat{\mu}_2, \widehat{C}_{11}(\boldsymbol{0})$ and $\widehat{C}_{22}(\boldsymbol{0})$, are the maximum likelihood (ML) estimates of μ_1, μ_2, $C_{11}(\boldsymbol{0})$ and $C_{22}(\boldsymbol{0})$, respectively.

The asymptotic properties of an estimator as in (4) have been studied in the literature for specific cases. Moreno et al. [2] studied the asymptotic properties of the ML estimator for a separable Matérn covariance model. They used a result provided by Mardia and Marshall [9] in an increasing domain sampling framework. Using this result and the delta method we can establish the following result for the Wendland-Gneiting model.

Theorem 1. Let $(Z_1(\boldsymbol{s}), Z_2(\boldsymbol{s}))^{\top}$, $s \in D$ be a bivariate Gaussian spatial process with mean $\boldsymbol{0}$ and covariance function given by

$$\boldsymbol{C}(\boldsymbol{h}) = \left[\rho_{ij}\sigma_{ii}\sigma_{jj}\left(1 + (\nu+1)\frac{\|\boldsymbol{h}\|}{b_{12}}\right)\left(1 - \frac{\|\boldsymbol{h}\|}{b_{12}}\right)_+^{\nu+1}\right]_{i,j=1}^2,$$

for $\nu > 0$ fixed. Define $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \rho_{12}, b_{12})^\top$ and denote $\widehat{\boldsymbol{\theta}}_n$ the ML estimator of $\boldsymbol{\theta}$. Then

$$\left(\nabla g(\boldsymbol{\theta})^\top \boldsymbol{F}_n(\boldsymbol{\theta})^{-1} \nabla g(\boldsymbol{\theta})\right)^{-1/2} (g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})) \xrightarrow{D} \mathcal{N}(0,1), \text{ as } n \to \infty,$$

where $g(\boldsymbol{\theta}) = \dfrac{2\rho_{12}\sigma_1\sigma_2 \left(1 + (\nu+1)\dfrac{\|\boldsymbol{h}\|}{b_{12}}\right)\left(1 - \dfrac{\|\boldsymbol{h}\|}{b_{12}}\right)_+^{\nu+1}}{\sigma_1^2 + \sigma_2^2}$, and $\boldsymbol{F}_n(\boldsymbol{\theta})$ is the Fisher information matrix associated with $\widehat{\boldsymbol{\theta}}$.

In addition, we also provide an expression for the asymptotic variance of the spatial concordance coefficient.

4 Applications

We explore the properties of the spatial concordance coefficient ρ^c for finite samples sizes through Monte Carlo simulations. Misspecification of the covariance function will also be addressed to inspect the impact on estimations when assuming a misspecified covariance function for the processes.

An application with real data also will be analyzed. Two images taken from the same location will be compared using the spatial concordance coefficient for different spatial lags. This measure of concordance is applied to digital images of forest-tree spring leaf-out obtained by different cameras. The result measures the agreement between two different acquisition processes. A concordance map similar to a codispersion map can be built to explore (an)isotropy in spatial concordance.

5 Acknowledgments

This work has been partially supported by the AC3E, UTFSM, under grant FB-0008, and by grants from the US National Science Foundation and NASA.

References

[1] Barnhart, H. X., Haber, M. J., Lin, L. I. (2007). An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics* **17**, 529–569.

[2] Bevilacqua, M., Vallejos, R., Velandia, D. (2015). Assessing the significance of the correlation between the components of a bivariate Gaussian random field. *Environmetrics* **26**, 545–556.

[3] Daley, D. J., Porcu, E., Bevilacqua, M. (2015). Classes of compactly supported covariance functions for multivariate random fields. *Stochastic Environmental Research and Risk Assessment* **29**, 1249–1263.

[4] Gneiting, T., 2002. Compactly supported correlation functions. *J*ournal of Multivariate Analysis **83**, 493–508.

[5] Hiriote, S., Chinchilli, V. M. (2011). Matrix-based Concordance Correlation Coefficient for Repeated Measures. *Biometrics* **67**, 1007–1016.

[6] Lin, L. I-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255–268.

[7] Lin, L., Hedayat, A. S., Sinha, B., Yang, M. (2002). Statistical methods in assessing agreement: Models, issues, and tools. *J*ournal of the American Statistical association **97**, 257–270.

[8] Lin, L., Hedayat, A. S., Wu, W. (2012). *Statistical Tools for Measuring Agreement*, Springer. New York.

[9] Mardia, K. V., Marshall, T. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135–146.

Stochastic Local Interaction Model for Spatial and Space-Time Data

Dionissios T. Hristopulos[1,*] and Vasiliki Agou[1,*]

[1] *School of Mineral Resources Engineering, Technical University of Crete, Chania, Greece 73100; dionisi@mred.tuc.gr*

Abstract. *Space-time data, even spatial data sometimes, are big. Thus, it is difficult to handle them by methods that scale poorly with size. The main roadblock in the application of geostatistical and machine learning methods (Gaussian processes) is the storage of dense covariance matrices and the $O(N^3)$ scaling of the numerical inversion of dense covariance matrices. Efficient representations of spatial or space-time correlations can be constructed using local dependence models, in the spirit of Gaussian Markov random fields (GMRFs). In the case of continuum random fields, the Gaussian field theories of statistical physics provide models with local structure. Stochastic local interaction (SLI) models are inspired both from GMRFs and Gaussian field theory. The main idea is that the correlations are generated by interactions between neighboring sites and times. The interactions are incorporated in a precision matrix with simple parametric dependence. The strength of the interactions and the size of the neighborhood are defined by means of kernel functions and respective bandwidths. Compactly supported kernels lead to finite-size local neighborhoods. This representation leads to sparse precision matrices. In addition, the precision matrix is explicitly constructed at the model estimation stage, which means that optimal prediction does not require the costly matrix inversion. Consequently, computational implementations require less memory space and run faster than traditional approaches. In the case of lattice data, SLI models transform into GMRFs. We present a specific SLI formulation and consider its application to lattice and scattered data.*

Keywords. *Spatial and spatio-temporal covariance modelling; Spatial/spatio-temporal lattice data*

1 Introduction

Space-time (ST) data are becoming available in overwhelming volumes and diverse forms due to the continuing growth of remote-sensing capabilities, the deployment of low-cost, ground-based sensor networks, as well as the increasing usage of sensors based on unmanned aerial vehicles, and crowdsourcing. The ongoing data explosion is expected to have an impact in various fields of science and engineering. The modeling and processing of massive datasets poses *conceptual, methodological, and technical* challenges. Sufficiently *flexible and computationally powerful* solutions that explicitly exploit ST structure are not widely available to date, because most existing methods are not designed for global, high-volume, hyper-dimensional, heterogeneous and uncertain ST data. For example, classical geostatistical and machine learning methods [1, 9] are limited by the cubic dependence of computational time on data size, which is prohibitive for large spatial data, even if time is omitted.

The modeling and processing of ST data are more complicated and computationally demanding than

purely spatial data. For example, theories that simply extend spatial statistics by adding a separable time dimension [2, 3] are often inadequate for capturing realistic correlations and for analyzing massive ST data. Current methods, whether they are based on geostatistics [1], spatio-temporal statistics [3], or machine learning [9] face serious scalability problems. A prevailing obstacle is the computationally demanding iterated inversion of *large covariance (Gram) matrices* [9, 11]. Hence, classical methods run on typical desktop computers are limited to datasets with size $N \sim O(10^3) - O(10^4)$. Approaches for alleviating the dimensionality problem (covariance tapering, composite likelihood, low-rank computations, stochastic partial differential equation representation) have been proposed and developed [11].

We present herein a framework for ST problems that is based on *stochastic local interaction* (SLI) models [4, 5]. For example, this formulation could be useful for filling gaps in ST records of meteorological variables that are needed for evaluation of renewable energy potential at candidate sites [6]. The main idea is that the correlations are determined by means of *sparse precision matrices* that only involve couplings between neighbors (in the ST domain). This idea of local dependence is underlying Markov random fields [10] and statistical field theories [7]. However, the former Markov random fields are typically used for regular lattice data while field theories are continuum models. SLI models extend the idea of locality to potentially scattered ST data.

2 ST Model based on Stochastic Local Interactions

A *space-time random field (STRF)* $X(\mathbf{s},t;\omega) \in \mathbb{R}$ where $\mathbf{s},t \in \mathbb{R}^d \times \mathbb{R}$ and $\omega \in \Omega$ is defined as a mapping from the probability space (Ω, A, P) into the space of real numbers $\mathbb{R}$. For each ST coordinate $(\mathbf{s},t)$, $X(\mathbf{s},t;\omega)$ is a measurable function of ω [2], where ω is the state index. The states (realizations) of the random field $X(\mathbf{s},t;\omega)$ are functions $x(\mathbf{s},t)$ obtained for a specific ω.

We will consider STRFs with joint pdf defined by the *Boltzmann-Gibbs* exponential distribution

$$f_{\mathrm{x}}[x(\mathbf{s},t)] = \frac{\mathrm{e}^{-\mathcal{H}[x(\mathbf{s},t)]}}{Z}, \tag{1}$$

where $\mathcal{H}[x(\mathbf{s},t)]$ is an energy function and Z is the normalizing factor known as partition function. Herein we will assume that the following properties are satisfied by $\mathcal{H}[x(\mathbf{s},t)]$ for any vector $\mathbf{x} = (x_1,\ldots,x_N)^\top$, where $N \in \mathbb{N}$ which comprises the field values at the ST point set $\{(\mathbf{s}_1,t_1)\ldots(\mathbf{s}_N,t_N)\}$:

- $\mathcal{H}[\mathbf{x}]$ is a quadratic function that can be expressed as

$$\mathcal{H}[\mathbf{x}] = \frac{1}{2}(\mathbf{x}-\mathbf{m}_{\mathrm{x}})^\top \mathbf{J}(\mathbf{x}-\mathbf{m}_{\mathrm{x}}). \tag{2}$$

- $\mathcal{H}[\mathbf{x}] > 0$ for all $\mathbf{x}$ that are not identically equal to zero. This is equivalent to the *precision matrix* $\mathbf{J}$ being a positive-definite matrix.
- The precision matrix $\mathbf{J}$ is a *sparse matrix* that embodies the local interactions.

More precisely, we are going to be concerned with the following SLI energy function

$$\mathcal{H}[\mathbf{x};\boldsymbol{\theta}] = \frac{1}{2\lambda}\left[\sum_{n=1}^{N}\frac{1}{N}(x_n - m_{\mathrm{x}})^2 + c_1\left\langle (x_n - x_k)^2 \right\rangle\right], \tag{3}$$

where $\boldsymbol{\theta}$ is the SLI parameter vector that includes the parameters m_x (mean value), λ (overall scale parameter proportional to the variance), and c_1 (dimensionless factor that determines the contribution from the squares of the increments $x_n - x_m$). The vector $\boldsymbol{\theta}$ includes additional parameters that determine the local ST neighborhood in the average $\langle\cdot\rangle$. The latter is defined by means of the *Nadaraya-Watson* average [8, 12], i.e.,

$$\langle (x_n - x_k)^2 \rangle = \frac{\sum_{n=1}^{N} \sum_{k=1}^{N} w_{n,k} \, (x_n - x_k)^2}{\sum_{n=1}^{N} \sum_{k=1}^{N} w_{n,k}}. \tag{4}$$

The weights $w_{n,k}$ are determined by means of *compactly supported kernel functions* $K(\cdot)$. Kernel functions are symmetric, $K(x) = K(-x)$, functions that take values in $[0,1]$. The argument on the kernel functions depends on the local neighborhood structure.

1. For example, we can use a *separable space-time neighborhood* so that $w_{n,k} = K(\mathbf{r}_{n,k}) \, K(\tau_{n,k})$ where $\mathbf{r}_{n,k} = (\mathbf{s}_n - \mathbf{s}_k)/h_n$ is the non-dimensional spatial lag between the initial point $\mathbf{s}_n$ and the target point $\mathbf{s}_k$, and $\tau_{n,k} = (t_n - t_k)/\alpha_n$ is the non-dimensional temporal lag between the initial and target times respectively. The spatial bandwidths h_n and the temporal bandwidths α_n are determined based on the geometry of the sampling network around the initial ST point $(\mathbf{s}_n, t_n)$. Note that this implies *asymmetric weights*, i.e, in general $w_{n,k} \neq w_{k,n}$.
2. Composite space-time metric:

$$\mathbf{J} = \frac{1}{\lambda} \left\{ \frac{\mathbf{I}_N}{N} + c_1 \, \mathbf{J}_1(\mathbf{h}) \right\}, \tag{5}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix: $[\mathbf{I}_N]_{i,j} = 1$ if $i = j$ and $[\mathbf{I}_N]_{i,j} = 0$ otherwise. The matrix $\mathbf{J}(\mathbf{h})$ is determined by the sampling pattern, the kernel function, and the bandwidths according to

$$[\mathbf{J}(\mathbf{h})]_{i,j} = -u_{i,j} - u_{j,i} + [\mathbf{I}_N]_{i,j} \sum_{l=1}^{N} [u_{i,l} + u_{l,i}], \tag{6}$$

$$u_{i,j} = \frac{K\left(\frac{\mathbf{s}_i - \mathbf{s}_j}{h_i}\right)}{\sum_{i=1}^{N} \sum_{j=1}^{N} K\left(\frac{\mathbf{s}_i - \mathbf{s}_j}{h_i}\right)}, \text{ where } i, j = 1, \ldots, N. \tag{7}$$

3 Conclusions

We have presented a framework for the construction of ST models that are based on the idea of local interactions. The model is based on exponential Boltzmann-Gibbs joint probability density functions. The local interactions in the above model are implemented by means of compactly supported kernel functions that compensate for the lack of a structured lattice. The model proposed herein represents Gaussian random fields with sparse precision matrices. Explicit expressions are given for ST prediction and for the estimation of the conditional variance. The sparse precision matrix representation leads to computational efficiency.

The formulation presented can be extended to multivariate random fields and to different local interaction models. In addition, it is possible to include anisotropic spatial distance metrics in the kernel functions, periodicity (in space and in time) by adding shifted averaged squared increments, and non-stationarity by allowing spatially dependent coefficients λ and c_1. These extensions will add flexibility to the model but reduce somewhat its computational efficiency.

Acknowledgments. This work was funded by the Operational Program "Competitiveness, Entrepreneurship and Innovation 2014-2020" (co-funded by the European Regional Development Fund) and managed by the General Secretariat of Research and Technology, Ministry of Education, Research, and Religious Affairs under the project DES2iRES (T3EPA-00017) of the ERAnet, ERANETMED_NEXUS-14-049. This support is gratefully acknowledged.

References

[1] J. P. Chilès and P. Delfiner (2012). *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York, 2nd ed.

[2] G. Christakos (1992). *Random Field Models in Earth Sciences*. Academic Press, San Diego.

[3] N. Cressie and C. L. Wikle (2011). *Statistics for Spatio-temporal Data*. John Wiley and Sons, New York.

[4] Hristopulos, D. T. (2015). Stochastic Local Interaction (SLI) model: Bridging Machine Learning and Geostatistics. *Computers and Geosciences* **85**(Part B), 26–37.

[5] Hristopulos, D. T. and Tsantili, I. (2017). Space-time covariance functions based on linear response theory and the turning bands method. *Spatial Statistics*, **22**(2), 321–337.

[6] Koutroulis, E. and Kolokotsa, D. (2010). Design optimization of desalination systems power-supplied by PV and W/G energy sources. *Desalination*, **258**(1-3), 171–181.

[7] Mussardo, G., 2010. *Statistical Field Theory: An Introduction to Exactly Solved Models in Statistical Physics*. Oxford University Press.

[8] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications* **9**(1), 141–142.

[9] C. E. Rasmussen and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press, Boston.

[10] H. Rue and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, Boca Raton, FL.

[11] Y. Sun, B. Li, and M. G. Genton (2012). Geostatistics for large datasets. In E. Porcu, J. M. Montero, and M. Schlather, editors, *Advances and Challenges in Space-time Modelling of Natural Events*, Lecture Notes in Statistics, pages 55–77. Springer Berlin Heidelberg.

[12] Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A*, **26**(1), 359–372.

Statistical post-processing of sea surface temperature forecasts

C. Heinrich[1,*], K. Hellton[1], A. Lenkoski[1] and T. Thorarinsdottir[1]

[1] *Norsk Regnesentral, Gaustadalleen 23a, Kristen Nygaards hus, NO-0373 Oslo; claudio.heinrich@nr.no, kristoffer.herland.hellton@nr.no, alex.lenkoski@nr.no, thordis.thorarinsdottir@nr.no*
[*] *Corresponding author*

Abstract. *Numerical weather prediction (NWP) models predict future weather by approximating solutions to the (deterministic) partial differential equations that govern the dynamics in atmosphere and oceans. These models oftentimes exhibit bias and miscalibration and therefore require statistical post-processing based on training data.*
We consider NWP forecasts for sea surface temperature on the entire globe issued by the Norwegian Climate Prediction Model NorCPM. Challenges for statistical post-processing of sea surface temperature are, among others, strong seasonality effects, trends in the bias caused by global warming, and a non-stationary spatial error correlation. Moreover, as we consider a fine grid spanning the entire globe, the dimension of the forecast variable is much higher than the sample size. In order to overcome these issues we apply principal component analysis to regularize the covariance matrix of the forecasting distribution.

Keywords. *Computational statistics; Geostatistics; Spatial and spatio-temporal covariance modelling; Weather and climate.*

---

1 Introduction

Numerical weather prediction (NWP) models are state of the art in modern meteorological forecasting. They rely on partial differential equations that describe the physics of the atmosphere and the oceans. Approximating a solution to these equations is used to obtain a prediction of the future weather from current observations. However, observations are not everywhere available and often imperfect and there are interactions in the atmosphere and oceans that are too complicated to model. As a consequence, even though they are very successful in capturing the dynamics of weather phenomena, NWP models are known to exhibit bias and misspecify forecast uncertainty and require statistical post-processing.

The goal of statistical post-processing is to correct a forecast model by observing the performance of the model over a training period, for which both observations and forecasts are available. Given a sufficient amount of training data, systematic errors of the forecast model can be found and then corrected, leading to a better forecast. It has been argued by [**?**] that *'the goal of probabilistic forecasting is to maximize the sharpness of the predictive distributions subject to calibration'*. A forecast is calibrated if the observations look like random draws from the forecasting distribution, whereas a sharper forecasting distribution has a lower spread and is therefore more informative.

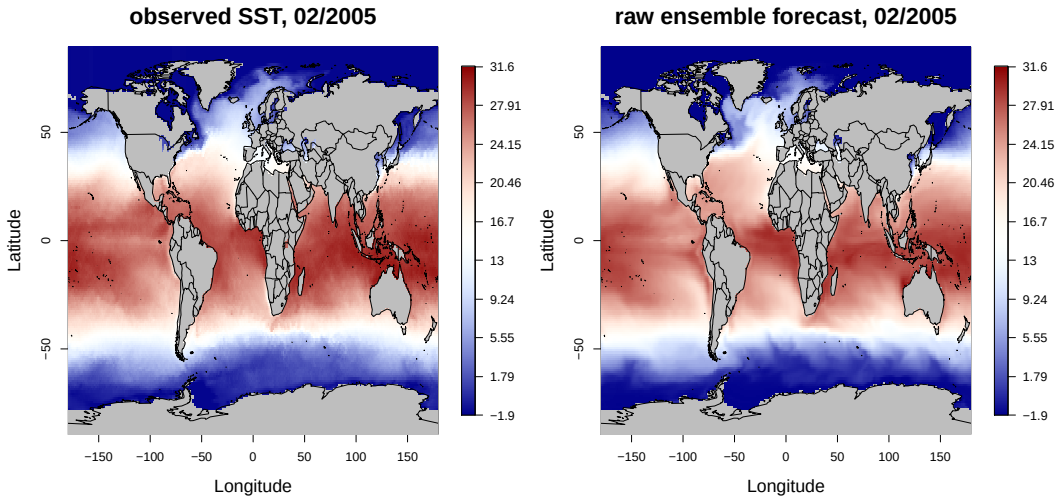


Figure 1: The left hand side plot shows the observed monthly mean sea SST for February 2005. The right hand side plot shows one member of the ensemble forecast for the same month that was issued in January 2005.

We consider monthly mean sea surface temperature (SST) predictions issued by the NWP model NorCPM (Norwegian Climate Prediction Model). The post-processing of the forecasts is divided into two steps, both of which together ensure that the forecast is calibrated and sharp: Bias correction and covariance modelling. Especially covariance modelling is challenging on a global scale, as the forecast error covariance is not stationary or isotropic in space, which is often a standard assumption in post-processing approaches of spatial phenomena. We rely on principal component analysis (PCA) to model the spatial covariance and demonstrate its favorable performance compared to other post-processing techniques. To the best of our knowledge, applying PCA in the context of statistical post-processing of forecasts is a novel approach.

2 A Model for the Forecast Residuals

The dataset available to us contains forecasts and observations of monthly mean SST on a grid with approximately 42,000 grid points spanning the entire globe for the years 1985–2010. The NorCPM issues a forecast ensemble with 9 members with new forecast runs being started every three months, consequently the lead time of the forecast is between one and three months. We use the years 1985-2000 as training data set and validate our methods on the years 2001-2010. Figure **??** shows an example for observed sea surface temperature and a member of the corresponding forecast ensemble. In this example, which is quite typical, the forecast is overall colder and has smoother spatial structures than the observation. Corresponding to these issues, our post-processing procedure is divided into two steps. First we estimate the bias of the forecast and correct for it. Then, we add noise to the bias corrected forecast that represents our forecast uncertainty and corrects the spatial covariance structure of the forecast. We follow the idea of nonhomogeneous Gaussian regression introduced by [**?**] in that we assume a Gaussian model for the residuals

$$\mathrm{res}_{y,m,s,k} := E_{y,m,s,k} - SST_{y,m,s},$$

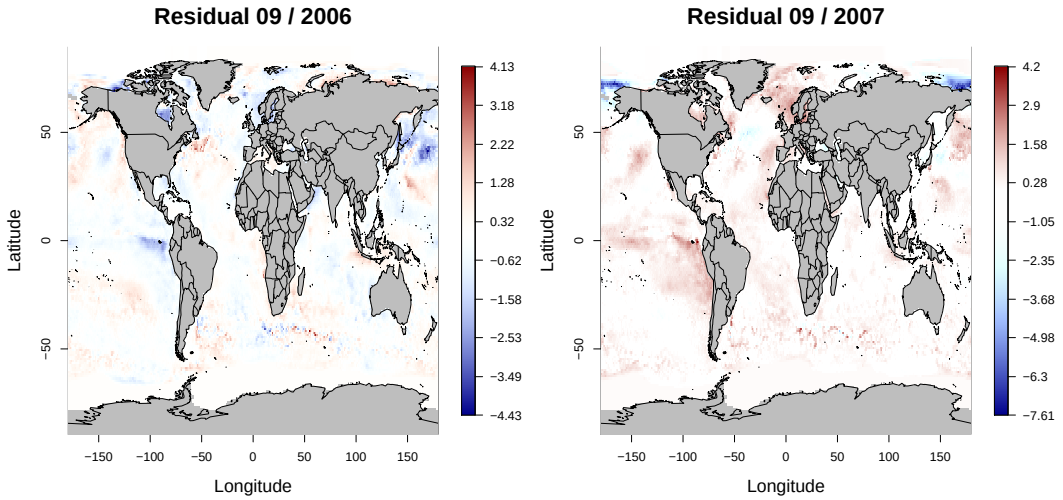


Figure 2: Example forecast residuals for September.

where *SST* denotes observed sea surface temperature and E denotes the ensemble forecast issued by NorCPM. The indices y, m, s and k represent the year, month, spatial location, and the ensemble member, respectively. Both the bias of the raw forecast ensemble and the forecast uncertainty depend strongly on the month of the year and the location. Thus, we model the marginal distribution of the residuals as

$$\text{res}_{y,m,s,k} \sim \mathcal{N}(b_{m,s}, \sigma^2_{m,s}),$$

where $b_{m,s}$ and $\sigma^2_{m,s}$ denote the bias and marginal variance for month m and location s. Both can be estimated from past data in a straightforward manner, for example by moving averages over past residuals (squared residuals) for the respective month and location. The main challenge lies in finding a good approximation of the spatial covariance structure,

$$\mathbf{res}_{y,m,k} - \mathbf{b}_m \sim \mathcal{N}_S(0, \Sigma_m),$$

where the vectors on the left hand side are S-dimensional, S being the number of grid points in our spatial grid.

3 Covariance Estimation

In Figure **??** we show two typical residuals for the month of September. They indicate that, additionally to positive correlation among grid points that are spatially close, we need to take into account covariance effects that correspond to ocean currents and other area effects such as islands and coastal regions. These effects are difficult to model explicitly and imply a non-stationary and anisotropic spatial covariance. Moreover, the grid size $S \approx 42{,}000$ is too large to use the full sample covariance matrix to estimate the covariance matrix Σ_m. Therefore, we apply principal component analysis in order to regularize the sample covariance matrix and reduce dimension. More precisely, we consider the sample covariance matrix

$$\widehat{\Sigma}_m = \frac{1}{KY-1} \sum_{k,y} \mathbf{res}_{y,m,k} \mathbf{res}^T_{y,m,k},$$

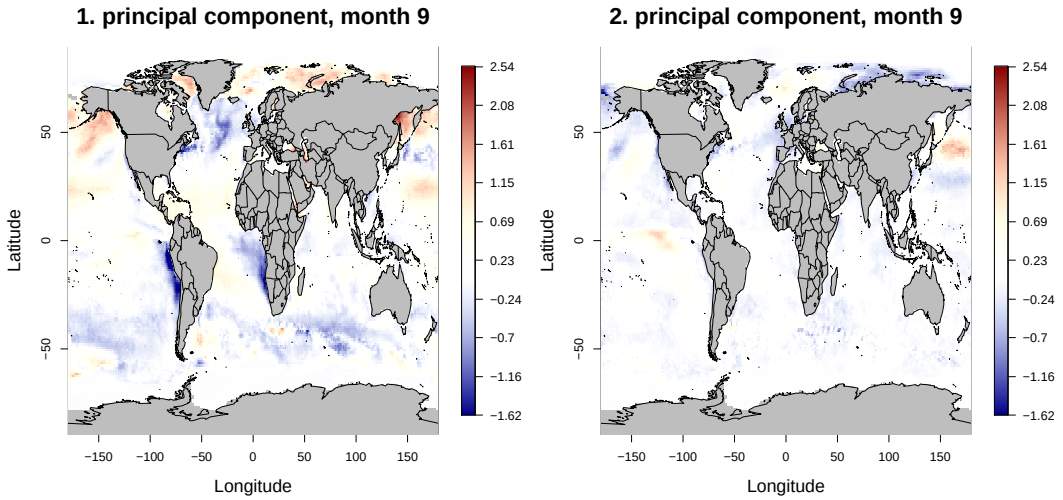


Figure 3: The first two principal components $\sqrt{\lambda_1}\mathbf{v}_1$ and $\sqrt{\lambda_1}\mathbf{v}_2$ for September.

where K is the ensemble size and Y is the number of years. Computing its eigenvalue decomposition

$$\widehat{\Sigma}_m = \sum_{i=1}^{KY} \lambda_i \mathbf{v}_i \mathbf{v}_i^T,$$

where $\lambda_1 > ... > \lambda_{KY}$ denote the eigenvectors of $\widehat{\Sigma}_m$ and $\mathbf{v}_i$ are orthonormal eigenvectors, we then estimate the covariance matrix as

$$\widehat{\Sigma}_m(d) := \sum_{i=1}^{d} \lambda_i \mathbf{v}_i \mathbf{v}_i^T,$$

where the cutoff level $d \ll KY$ is a tuning parameter for our estimate. Figure **??** shows the first two (scaled) eigenvectors $\sqrt{\lambda_1}\mathbf{v}_1$ and $\sqrt{\lambda_1}\mathbf{v}_2$ for the month of September, given visual evidence that the PCA is able to uncover much of the covariance structure that seems to be present in Figure **??**.

We underline this first impression by applying thorough checks for calibration and comparing our method to other competing methods, such as ensemble copula coupling and geostationary models. The superior performance of the principal component method along with its high interpretability and the dimension reduction, which allows for fast sampling from the forecasting distribution, lets us believe that PCA has a wide potential in post-processing of high dimensional forecasts.

References

[1] Gneiting, T., Balabdaoui, F. and Raftery, A.E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* **69**, 243–268.

[2] Gneiting, T., Raftery, A.E., Westveld, A.H. and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.* **133**, 1098–1118.

Disaggregation of large-scale atmospheric data: a non-deterministic geostatistically-based approach

S. Chen[1], E. Leblois[1,*], S. Anquetin[2] and S. Martino[3]

[1] *sheng.chen@irstea.fr, etienne.leblois@irstea.fr*
[2] *sandrine.anquetin@univ-grenoble-alpes.fr*
[3] *sara.martino@sintef.no*
[*] *Corresponding author*

Abstract. *In this contribution, a geostatistically sound approach inspired by a soil simulation strategy known as the "pilot point" technique is proposed to simulate heterogeneous spatial fields respecting average values known over large scale domains. It typically allows the disaggregation of atmospheric reanalysis, GCMs outputs or large scale stochastic weather simulations. The disaggregation is based on an "a priori" small-scale simulation over the final grid, conducted as in [4]. The two main components of this "a priori" simulation are a first Gaussian field thresholded to get a 0/1 rainfall indicator field and a second Gaussian field to get a non-zero rainfall field, the non-zero rainfall field is then multiplied by the indicator field. The novelty is that the Gaussian fields are iteratively modified so that the final simulation will reach the wished large-scale values. The disaggregation is not deterministic and reintroduces small-scale variability implicit in the large-scale data, giving an instrumental picture of the conditional uncertainty. The technique was developed within a stochastic weather simulation project led by Sintef; it is presently tested under Norwegian and French climates.*

Keywords. *Geostatistics; Weather and Climate.*

---

1 Introduction

In hydrology and renewable energy projects, one may need to disaggregate values available as large scale simulation (LS-values). A typical case is the one of rainfall quantities. There is not one only solution to this question: a conditional variability in the fine grid is natural and should be generated by the disaggregation technique. Our contribution describes a technique to disaggregate precipitation quantities known on control domains into cells of a fine grid, trying to keep a realistic distribution supposedly known at the small scale in terms of local marginal and spatio-temporal structure. The technique then build ensembles where all members are equally probable under the assumptions. By design, the technique strictly respects the values given at LS. Any possible errors in the large-scale input values must be cleaned up before disaggregation.

2 Method

2.1 Inspiration

The technique is inspired by the solution for solving inverse problems in hydrogeology given by [2] under the name of “pilot points approach”. Their motivation was aquifer reconstruction given observed macro-scale properties. Adding arbitrary control points with adjustable values, [2] were able to tune the aquifer as wished. Here control values are defined on the LS-domains, not on points, hence "pilot values approach" naming is more relevant in our case.

2.2 Algorithm

Two steps are achieved in the disaggregation method. First a small-scale intermittency field is generated and needs to be adjusted to respect the LS value. Second a small-scale non-zero rainfall field is generated and multiplied by the intermittency field (to delineate actually rainy areas). The final composite needs to respect the LS non-zero precipitation value.

More precisely, the intermittency simulation step is as follows and will be latter illustrated in the next section: 1/ generate a Gaussian field for the small-scale grid having the expected spatio-temporal structure suitable for thresholding-based simulation of intermittency; 2/ check how the thresholded field compares with the expected large-scale wetness values; 3/ where the recovered wetness is too much, gently lower the Gaussian field ; where the recovered wetness is not enough, gently higher the Gaussian field. Adjust as necessary to recover every prescribed large-scale wetness value.

Then, the non-zero rainfall step is presented as follows: 1/ generate a Gaussian field for the small-scale grid having the spatio-temporal structure suitable for anamorphosis-based simulation of non-zero rainfall; 2/ check how the anamorphosed field compares with the expected large-scale rainfall values (keep only the average on wet cells); 3/ where the recovered precipitation amount is too much, gently lower the Gaussian field; where the recovered precipitation amount is not enough, gently higher the Gaussian field. Adjust as necessary to recover every prescribed large-scale wetness value.

How to “gently change” a Gaussian field ?
The idea is to choose one scalar shift value per control zone (LS cell); using block-to-point kriging [3], these shifts are distributed (interpolated) to the small-scale grid and this distributed shift is added to the Gaussian grid. (A warning : what is easily found about block kriging usually refers to point-to-block kriging, where the kriging matrix is between data points and the target is a domain of finite non point size, so a block. Here we really mean block-to-points kriging, where the kriging matrix is build on block covariance between data blocks and the right vector is covariance between data blocks and target point. As the kriging variance is not needed, we recommend to use dual kriging [6].

How to respect the exact values of large scale ?
The previous paragraph has explained briefly how we “gently change ” a Gaussian field, so that the bloc in the Gaussian field will respect the given value over the large scale blocs. But because of the non-linear transformation to the user field (intermittency) it is not obvious how to choose the bloc values to condition the underlying Gaussian field. The dichotomy method is a root-finding method that repeatedly bisects an interval and then selects a sub-interval in which a root must lie for further processing [1, 2]. In

our context, the dichotomy method will allow each control zone of an arbitrary Gaussian field to come close to prescribed large-scale value by gently changing the Gaussian field in each iteration, until the large-scale values are respected for all control zones.

The suggested technique makes a nice job in control runs. To achieve acceptable performance, a careful planification of the basic spatio-temporal integrals involved is recommended. We are aware that a link between intermittency and non-zero rainfall values [5] may exist. In our case, it can be present in the large scale values, then it will respected. It is left for further studies to possibly introduce it in the disaggregation also.

3 Application: the Cévennes-Vivarais region, France

Figure 1 presents the control zones (LS cells) delineation and the small-scale 2*km* resolution grid. The LS cells (i.e. the homogeneous rainfall zone) are described by the two known values, the average daily precipitation (i.e. average rain over the whole LS cell) and the daily rainfall intermittency (i.e. fraction of grid cells in the LS cell presenting a non-zero rain). As an illustration, Table 1 presents 3 continuous daily simulated values, obtained with the copula based parametric model that are used as input in the disaggregation model. Figure 2 presents the simulation steps of the proposed disaggregation technique for an illustrative sequence of 3 successive days.

4 Tables and figures

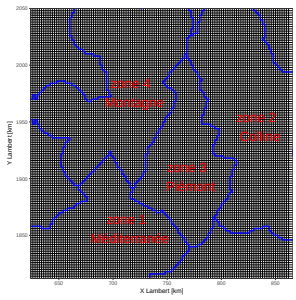


Figure 1: Gridded Cévennes-Vivarais region. Large-scale control domain (blue lines) and small-scale regular grid (black squares) at 2 km resolution.

t	P_1	I_1	P_2	I_2	P_3	I_3	P_4	I_4
t_1	2.4	0.6	3.5	0.4	1.5	0.8	5.5	0.8
t_2	0	0	3.7	0.9	0.8	0.4	4.5	0.9
t_3	0.3	0.3	0.7	0.1	0.1	0.2	1.5	0.7

Table 1: values to be respected for average precipitation (P) and the rainfall intermittency (I), for 3 successive days, in the 4 homogeneous zones.

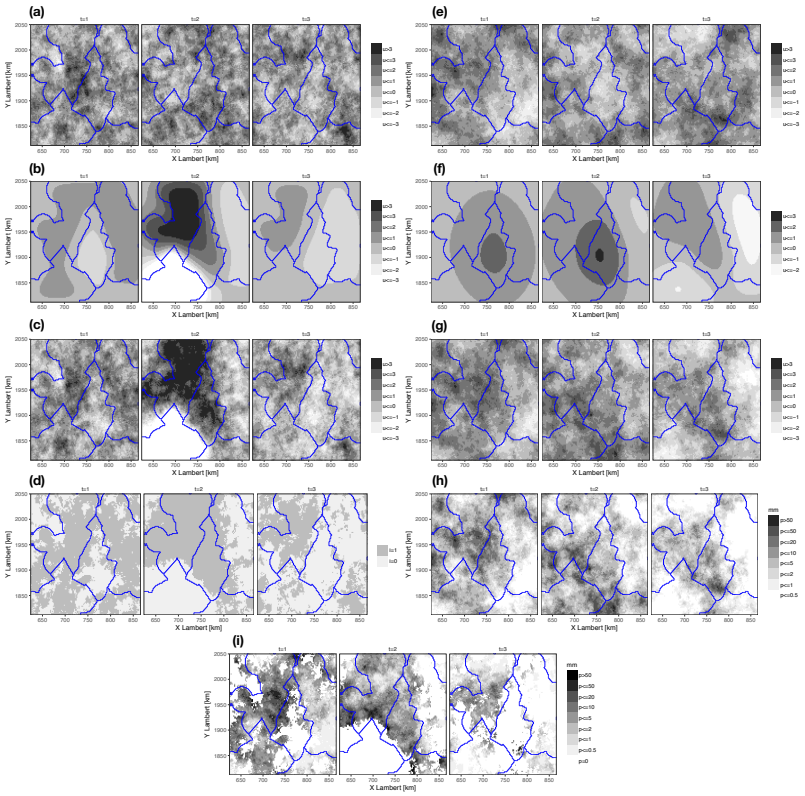


Figure 2: The simulation steps ; all fields are spatio-temporal (3 days, 4 zones) a) Gaussian field for intermittency simulation. b) Interpolated pilot values, to be added to the previous c) Sum of the two previous, to be tresholded for intermittency simulation d) Final simulated intermittency. d) Gaussian field for non-zero rainfall simulation f) Interpolated pilot values to be added to the previous g) The sum of the previous, to enter anamorphosis towards the user distribution h) Non-zero rainfall field to be intersected with intermittency i) Final rainfall field.

References

[1] Certes, C., and G. de Marsily (1991), Application of the pilot point method to the identification of aquifer transmissivities. *Advances in Water Resources* **14(5)**, 284–300.

[2] de Marsily, G., J.-P. Delhomme, F. Delay, and A. Buoro (1999), Regards sur 40 ans de problèmes inverses en hydrogéologie. *Comptes Rendus de l'Académie des Sciences-Series II A-Earth and Planetary Science* **329(2)**, 73–87.

[3] Kerry, R., P. Goovaerts, B. G. Rawlins, and B. P. Marchant (2012), Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale. *Geoderma* **170(Supplement C)**, 347–358.

[4] Leblois, E., and J.-D. Creutin (2013), Space-time simulation of intermittent rainfall with prescribed advection field: Adaptation of the turning band method, *Water Resources Research* **49(6)**, 3375–3387.

[5] Schleiss, M., Chamon, S. and Berne, A. (2014), Stochastic simulation of intermittent rainfall using the concept of "dry drift". *Water Resources Research* **50**, 2329–2349.

[6] Royer, J.-J., and P. Vieira (1984), Dual formalism of kriging. *Geostatistics for natural resources characterization* **2**, 691–702.

Two-Scale Spatial Models for Binary Data

C. Hardouin[1,*] and N. Cressie[2]

[1] *MODAL'X, Université Paris Nanterre, France; hardouin@parisnanterre.fr*
[2] *NIASRA, University of Wollongong, Australia; ncressie@uow.edu.au*
[*] *Corresponding author*

Abstract. *A spatial lattice model for binary data is constructed from two spatial scales linked through conditional probabilities. A coarse grid of lattice locations is specified and all remaining locations (which we call the background) capture fine-scale spatial dependence. Binary data on the coarse grid are modelled with an autologistic distribution, conditional on the binary process on the background. The background behaviour is captured through a hidden Gaussian process after a logit transformation on its Bernoulli success probabilities. The parameters of the new model come from both spatial scales, and are estimated with likelihood-based methods. We introduce the Spatial odds-ratio, which is more appropriate in the binary context than the spatial correlation. Presence-absence data of corn borers in the roots of corn plants are used to illustrate how the model is fitted.*

Keywords. *Gaussian process; Auto-logistic model; Spatial odds-ratio; Laplace approximation.*

---

1 Introduction

Binary spatial data are involved in various domains. One common model for regularly spaced binary data is the auto-logistic model, which belongs to Besag's auto-models class ([1]). In a hierarchical framework, a Generalized Linear Model ([4], [5]) can be implemented with a link appropriate for binary data. The logit link is canonical and a natural choice for the hidden process is Gaussian. In this work, we focus on binary data on a spatial lattice with two spatial scales. For the sake of simplicity, we assume that the process is observed on a regular lattice $D = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\} \subset \mathbb{R}^2$. Let $\mathbf{Z} = (Z(\mathbf{s}) : \mathbf{s} \in D)^{\mathrm{T}}$ be the process on D, taking its values in the state space $E = \{0,1\}^D$.

We consider two scales of spatial dependence, which occur locally at fine-scale resolution 1, and at a coarse-scale resolution $\Delta > 1$. This distance Δ is assumed known; in practice, it may be obtained from a preliminary exploratory analysis of the data, or by subject matter experts. We then specify a coarse regular grid of sites at resolution Δ; the locations on the coarse grid define what we call the *Grid*, and all remaining locations on the underlying lattice define what we call the *Background*.

The spatial-process modelling is the following. We write $\mathbf{Z} = (\mathbf{Z}_G^{\mathrm{T}}, \mathbf{Z}_B^{\mathrm{T}})$. Then we start with the Background, which involves a conditional logistic model for $\mathbf{Z}_B$; then, conditional on $\mathbf{Z}_B$, we define $\mathbf{Z}_G$ on the Grid.

2 Two-scale spatial modelling

2.1 Fine-scale process on the Background

We model the binary variables using a Bernoulli distribution, where the mean depends on an underlying zero-mean Gaussian spatial process ε with covariance matrix Σ. Moreover, we assume conditional independence of the Bernoulli random variables given ε. Thus we consider,

$$Z_B(\mathbf{s}) \mid \varepsilon(\mathbf{s}) \sim Ber(p(\mathbf{s})), \tag{1}$$

where $p(\mathbf{s}) = \dfrac{e^{\varepsilon(\mathbf{s})}}{1+e^{\varepsilon(\mathbf{s})}}$.

2.2 Coarse-scale process on the Grid

We define the model on the Grid conditional on the Background using a Markov Random Field model with a neighborhood graph on the Grid, which recall has resolution Δ. For the sake of simplicity we consider here the four nearest neighbours, but the model can be modified easily to account for extra spatial dependence.

For each site $\mathbf{s} \in G$, we define the four-nearest-neighbourhood set $N_G(\mathbf{s}) = \{\mathbf{u} \in G : \mathbf{u} = \mathbf{s} \pm (\Delta, 0), \mathbf{s} \pm (0, \Delta)\}$. Our conditional model for the Grid values is:

$$\pi_{\mathbf{s}}(Z_G(\mathbf{s}) \mid \mathbf{Z}_B, \mathbf{Z}_{N_G(\mathbf{s})}) = \frac{\exp\left\{\alpha_B(\mathbf{s}) Z_G(\mathbf{s}) + \frac{\beta}{4} \sum\limits_{\mathbf{u} \in N_G(\mathbf{s})} Z_G(\mathbf{s}) Z_G(\mathbf{u})\right\}}{1 + \exp\left\{\frac{\beta}{4} \sum\limits_{\mathbf{u} \in N_G(\mathbf{s})} Z_G(\mathbf{s}) Z_G(\mathbf{u})\right\}}, \tag{2}$$

Here, β is the spatial interaction parameter, which we assume to be constant over the Grid; $\alpha_B(\mathbf{s})$ captures the dependence on $\mathbf{Z}_B$, with

$$\alpha_B(\mathbf{s}) = \gamma + \alpha \times \frac{\sum_{\mathbf{u} \in N_B(\mathbf{s})} Z_B(\mathbf{u})}{|N_B(\mathbf{s})|},$$

where $N_B(\mathbf{s}) = \{\mathbf{u} = (u_1, u_2) \in B : \ |u_1 - s_1| \leq \Delta, \ |u_2 - s_2| \leq \Delta\}$ is the set of the neighbour Background locations. We ultimately choose $\gamma = -\dfrac{\alpha + \beta}{2}$.

3 Simulation experiments

We show through simulations that the two-scale spatial model presented before allows both competitive or cooperative behaviours as well. Further, we introduce the spatial odds ratio, which accounts for dependence better than the spatial correlation when the data are binary.

4 Estimation

The joint distribution of the process is the product of two terms, corresponding to the distribution on the Grid given the Background, times the distribution on the Background. Due to the intractable normalizing constant, the auto-model parameters are estimated by maximizing the pseudo-likelihood introduced by Besag [2]. The second term involves the latent Gaussian field's parameters; their estimation in a hierarchical statistical model typically requires an EM algorithm; see [3] or [7]. The E-step needs the expectation of the latent field ε given the observations, but we do not know the integrated distribution. Here we use Laplace approximations to approximate the intractable integrals.

5 Application to Corn Borers dataset

An extensive entomological field study of European corn borer larvae was conducted in northwest Iowa ([6]). We consider a square area divided into 324 regular subplots, on 18 rows, planted with corn seeds. The response variables analyzed were defined as binary variables for the subplots, where the value 0 was obtained if corn borer larvae were absent, and the value 1 was obtained if one or more larvae were present. First, the resolution Δ of the Grid, and the location of the Grid, are chosen according to a preliminary exploratory step, by inspecting the spatial odds ratio at different lags. Then, we estimate the parameters of the model.

References

[1] Besag J., 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society, Series B (Methodological), Vol. 36, No. 2., pp.192-236.

[2] Besag J., 1977. Efficiency of pseudo likelihood estimation for simple Gaussian fields. Biometrika 64, pp. 616–618.

[3] Dempster A.P., Laird N., Rubin D., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39, pp. 1-38.

[4] McCullagh P., Nelder J.A., 1989. Generalized Linear Models, second ed. Chapman and Hall, London, UK.

[5] McCulloch C.E., Searle S.R., Neuhaus J.M., 2001. Generalized, Linear, and Mixed Models. Wiley, New York, NY.

[6] McGuire J., Brindley T., Bancroft T., 1957. The distribution of European corn borer larvae Pyrausta nubi/alis (Hbn.) in field corn, Biometrics 13, pp. 65-78.

[7] McLachlan G.J., Krishnan T., 2008. The EM Algorithm and Extensions, second ed. Wiley-Interscience, New York, NY.

Comparing two models for disease mapping data not varying systematically in space

Helena Baptista[1,*], Jorge M. Mendes[1], Peter Congdon[2]

[1] *NOVA Information Management School, Universidade Nova de Lisboa, Lisboa, Portugal*
[2] *School of Geography and Life Sciences Institute, Queen Mary, University of London, UK*
[*] *NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal, mhbaptista@novaims.uml.pt*

Abstract. *Conditionally specified Gaussian Markov random field (GMRF) models with adjacency-based neighborhood weight matrix, commonly known as neighborhood-based GMRF models, have been the mainstream approach to spatial smoothing in Bayesian disease mapping. However, there are cases when there is no evidence of positive spatial correlation or the appropriate mix between local and global smoothing is not constant across the region being study. Two models have been proposed for those cases, a conditionally specified Gaussian random field (GRF) model using a similarity-based non-spatial weight matrix to facilitate non-spatial smoothing in Bayesian disease mapping, and a spatially adaptive conditional autoregressive prior model. The former model, named similarity-based GRF, is motivated for modeling disease mapping data in situations where the underlying small area relative risks and the associated determinant factors do not varying systematically in space, and the similarity is defined by similarity with respect to the associated disease determinant factors. In the presence of disease data with no evidence of positive spatial correlation, a simulation study showed a consistent gain in efficiency from the similarity-based GRF, compared with the adjacency-based GMRF with the determinant risk factors as covariate. The latter model considers a spatially adaptive extension of Leroux et al. [9] prior to reflect the fact that the appropriate mix between local and global smoothing may not be constant across the region being studied. Local smoothing will not be indicated when an area is disparate from its neighbours (e.g. in terms of social or environmental risk factors for the health outcome being considered). The prior for varying spatial correlation parameters may be based on a regression structure which includes possible observed sources of disparity between neighbours. We will compare the results of the two models.*

Keywords. *Bayesian modelling. Disease mapping. Small area.*

---

1 Introduction

Spatial disease mapping models are being extensively used to describe geographical patterns of mortality and morbidity rates. Information provided by these models is considered invaluable by health researchers and policy-makers as it allows, for example, to effectively allocate funds in high risk areas, and/or to plan for localised prevention/intervention programmes.

In cases of rare diseases and/or low populated areas, the classical estimators of the morbidity rates

show high variability, and spatial disease mapping models overcome that by borrowing strength from spatial *neighbours*. One rationale is that the spatial random effects used to implement such borrowing of strength are proxies for unobserved risk factors that vary smoothly in space. Models used in disease mapping are usually generalized linear mixed models (GLMM) formulated within a hierarchical Bayesian framework, and Poisson likelihood is often assumed for data in the form of counts of cases for each areal unit. Neighbourhood information is explicitly incorporated into the model by means of an appropriate prior specification. The seminal work of Besag et al. [3] provides a pair of area-specific random effects to model unstructured heterogeneity (extra-Poisson variation) and spatial similarity. The Besag-York-Mollié (BYM) model is an extension of the intrinsic conditional autocorrelation (CAR) model, a well known Gaussian Markov random field (GMRF) prior in disease mapping [3]. In the same field, Leroux et al. [9] proposed a conditional autoregressive prior incorporating a spatial correlation parameter, with its extreme values corresponding to pure spatial and pure unstructured residual variation. One important aspect of the CAR modelling is the definition of the so-called neighbourhood matrix, which characterizes the spatial structure of the data at hand, and is based on the concept of *neighbours*.

The debate on the definition of *neighbours* can be traced back to Besag [2]. Others have worked in defining *neighbours* in several different ways (Besag et al. [3], Best et al. [4], Earnest et al. [6], Congdon [5] and Lee and Mitchell [8]).

Most of the research in disease mapping is related with diseases resulting from environmental exposures, such as respiratory complications and cancer. Those extrinsic disease determinant factors are spatially smoothed, and using some kind of spatial proximity, either by adjacency or by distance, between areas in the definition of *neighbours* has therefore provided good results. In cases in which no spatial positive autocorrelation is displayed by the data, the neighbourhood matrix as it exists today may not be adequate. The similarity-based GRF approach, proposed by Baptista et al. [1], replaces the neighbourhood-based GMRF approach. The structure of the conditionals is maintained, but the smoothing and borrowing strength mechanisms are now based on the similarity of the areas, regardless of their relative location in space.

Another approach to the same aspect is proposed by Congdon [5], where is considered that uniform borrowing of strength based simply on proximity or contiguity may not be appropriate when there are discontinuities in the spatial pattern of health events or risk factors; for instance, a low mortality area surrounded by high mortality areas. Such discontinuity may often reflect spatial discontinuities in risk factors, whether observed or unobserved. An area showing such discontinuity may have a distorted smoothed rate when smoothing is towards the local mean.

In this submission we will present results of the implementation of the above two mentioned models. Section 2 will provide a brief overview of the similarity model, Section 3 will provide a brief overview of the adaptive model and Section 4 will provide a brief overview of the data used. This is still work in progress, so no results will be presented now.

2 A similarity-based Gaussian random field model

The GRF model proposed no longer retains the Markovian properties as those based on the neighbourhood weights. Instead of using spatial distance or spatial adjacency, a measure reflecting similarity between areas is introduced. Data used should come from: a) a disease determinant factor or a combination of factors, b) a source external to the survey that collected the disease data. The main objective of the

proposed model is the provision for borrowing strength between areas with similar disease determinant factors.

Firstly, regions exhibiting the *same or close* level of risk in a determinant factor will be regions with the *same or close* risk of the disease. Secondly, if disease data need to be *strengthened*, using disease determinant factor information collected by the same survey might inflate or not remediate possible *weaknesses* of the disease data. Therefore, an external source for the disease determinant factor is critical.

The rationale of our approach is the following: in cases of diseases with no environmental determinant factors, use of a positive spatial correlation based on physical distance or adjacency, in the GRF/GMRF model, may not be the best way to reflect similarity between areas. By using the GRF model reflecting *how similar* each area is to one another, in terms of a disease determinant factor that was collected by an external source, the disease risk distribution can be better assessed.

We use the BYM model (more detailed specifications can be found elsewhere [3]), with a neighbourhood matrix based on a matrix definition proposed by Best et al. [4], the new similarity matrix, with elements h_{ij} for each region i, with the following structure:

$$h_{ij} = \begin{cases} e^{-p_{ij}/b}, & \text{if } j \neq i \\ \frac{1}{n-1}\sum h_{(-i)}, & \text{otherwise,} \end{cases}$$

where p_{ij} is the absolute gap between region i and region j, $p_{ij} = |p_i - p_j|$, in terms of the disease determinant factor, and b is equal to a value that gives a relative weight of 1% ($h_{ij} = 0.01$) to an area i whose difference from an area j is the mean inter-region difference for the country. Elements h_{ii} need a specific definition, otherwise their value would be the one contributing the most to the prior, as $e^0 = 1$ and all other d_{ij} elements have values between 0 and 1. Therefore, p_{ii} values are equal to the average value of all elements except the ith area value.

More details can be found in Baptista et al. [1].

3 A Spatially Adaptive Conditional Autoregressive Prior

The similarity based GRF prior (Section 2) replaces spatial proximity as a basis for borrowing strength by similarity in one or more risk factors, and so takes explicit account of the actual spatial pattern of risk factors, allowing for the case when that pattern may be irregular (not spatially smooth). By contrast, the spatially adaptive approach retains the principle of spatial borrowing of strength, but modifies it to better represent discontinuities in the outcome and/or observed risk factors. The degree of spatial correlation is allowed to vary between sub-regions of the region under consideration, with one possible scheme linking varying spatial correlation to spatial similarity (or dissimilarity) in risk factors between an area and its surrounding locality.

We start with the Leroux et al. [9] model and here we propose spatial adaptivity based on area specific $\lambda \in [0,1]$, the uniform measure of spatial dependence. For areas i, distinctly low λ_i correspond to spatial disparate areas, unlike their neighbours in health risk and/or risk factors, so that there may be benefit in downweighting the principle of uniform pooling to the locality mean.

If predictors W_i measuring dissimilarity in observed risk factors are available, and so relevant to

whether pooling should be local or global, one can use a regression scheme $logit(\lambda_i) \sim N(W_i\gamma, 1/\tau_\lambda)$, where γ are regression parameters. For example, in Congdon (2008) the discrepancy measure is based on area socioeconomic deprivation z_i, with dissimilarity represented as $W_i = |z_i - \bar{Z}_i|$ with $\bar{Z}_i$ being average deprivation in the locality L_i around area i, namely $\bar{Z}_i = \sum_{j \in L_i} z_j/d_i$.

More details can be found in Congdon [5].

4 The data

The relative merits of the methodologies mentioned in sections 3 and 4 must be investigated by applying those models to data sets exhibiting different patterns of spatial association. Therefore, both models are assessed under the spatial association generated by data sets used either in Baptista et al. [1] (alcohol abuse disorder) and Congdon [5].

Other data sets are under investigation and may be used.

References

[1] Baptista, H., Mendes, J. M., MacNab, Y. C., Xavier, M., & de Almeida, J. M. C. (2016). A Gaussian random field model for similarity-based smoothing in Bayesian disease mapping. *Statistical Methods in Medical Research, 25(4)*, 1166-1184.

[2] Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society, 36*, 192–236.

[3] Besag, J., York, J., & Mollié, A. (1991). Bayesian Image Restoration, with Two Applications in Spatial Statistics. *Annals of the Institute of Statistical Mathematics, 43*, 1–20.

[4] Best, N.G, Arnold, R.A., Thomas, A., Waller, L.A. & Conlon, E.M. (1999). Bayesian Models for Spatially Correlated Disease and Exposure Data. *Bayesian Statistics 6* , 131–147.

[5] Congdon, P. (2008). A spatially adaptive conditional autoregressive prior for area health data. *Statistical Methodology, 5*, 552–563.

[6] Earnest, A., Morgan, G., Mengersen, K., Ryan, L. ,Summerhayes, R., & Beard, J. (2007). Evaluating the effect of neighbourhood weight matrices on smoothing properties of Conditional Autoregressive (CAR) models. *International journal of health geographics, 6*, 54.

[7] Lee, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and spatio-temporal epidemiology, 2*, 79–89.

[8] Lee, D., & Mitchell, R. (2013). Locally adaptive spatial smoothing using conditional auto-regressive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 62*, 593–608.

[9] Leroux, B. G., Lei, X., & Breslow, N. (2000). Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence. In M. E. Halloran and D. Berry (Eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (Vol. 116, pp. 179-191). New York, NY: Springer New York.

[10] MacNab, Y. (2011). On Gaussian Markov random fields and Bayesian disease mapping. *Statistical methods in medical research, 20(1)*, 49–68

[11] Sun, D., Tsutakawa, R., & Speckman, P. L. (1999) Posterior distribution of hierarchical models using CAR distributions. *Biometrika, 86(2)*, 341-350.

Spatio-temporal models for georeferenced unemployment data

S. Pereira[1,*], K. F. Turkman[1], L. Correia[2] and H. Rue[3]

[1] *CEAUL - FCUL, Universidade de Lisboa, Lisboa, Portugal; soraia.gpereira@gmail.com, kfturkman@fc.ul.pt*
[2] *Instituto Nacional de Estatística, Lisboa, Portugal; luis.correia@ine.pt;*
[3] *King Abdullah University of Science and Technology, Saudi Arabia; haavard.rue@kaust.edu.sa;*
[*] *Corresponding author*

Abstract. *In Portugal, the Portuguese National Statistical Institute publishes quarterly labour market figures at national level and for both NUTS I and NUTS II regions. Over recent years it has become increasingly important to know these figures at more disaggregated levels. However, using the current estimation method, it is not possible to produce satisfactorily precise estimates. This problem is known in the literature as 'small area estimation'. Some alternative methods have subsequently been proposed, one of which - and perhaps the most important - is the Fay-Herriot model, an areal model which assumes normality of the data. However, the assumptions made in this model are very restrictive and do not appear to be suitable in the context of unemployment. From the 4th quarter of 2014 onwards, all the sampling units (the residential buildings) of the Portuguese Labour Force Survey (PLFS) were georeferenced. To take advantage of this, the authors proposed using this new data, along with the information regarding the inhabitants themselves. Thus, the method we propose is based on a point referenced data model, also described as a geostatistical model. This model assumes that the points in the population are fixed and the interest is to model the spatial variation of the marks. Here, the points are the residential buildings, whereas the associated marks are the number of unemployed people residing in these buildings. The inference will be based on the Integrated Nested Laplace approximations (INLA).*

Keywords. *point-referenced data models; Bayesian inference; unemployment estimation; small area estimation; INLA*

---

1 Introduction

In Portugal, the National Statistical Institute (NSI) is responsible for performing, on quarterly basis, the Labour Force Survey (LFS) covering the entire national territory and for supplying the national and European entities the conclusions taken from these sample surveys. Consequently, the NSI publishes official quarterly labour market statistics, including the estimated unemployment figures at different spatial resolutions, typically for NUTS I and NUTS II regions. NUTS is the classification of territorial units for statistics, created by the Eurostat and the National Statistical Institutes of European Union, and it includes three hierarchical levels: NUTS I, NUTS II and NUTS III (see figure 1).

Together with the increase in demand for ever detailed information at higher spatial resolutions, the demand for more reliable estimates without increasing the cost of larger samples also increases.

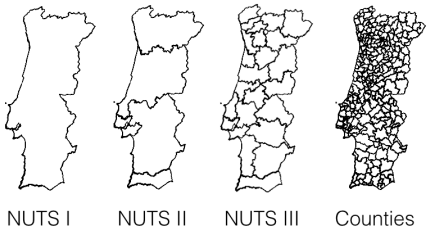


Figure 1: NUTS (version 2013) and counties in mainland Portugal

Typically, NSI produces unemployment estimates based on direct estimation methods based on Horvitz-Thompson estimator. However, these direct estimation methods do not perform well in small areas, increasing the demand either for larger samples or for small area estimation methods (Rao and Molina, 2015) that borrow strength from neighboring observations.

There have been considerable methodological developments to solve small area estimation problems in an unemployment context. The majority of small area methods are based on generalized linear models applied to areal data by modelling an appropriate counting process. These methods can "borrow strength" from area to area and make use of auxiliary information at regional level, compensating for the small sample sizes in each area due to the designed sampling survey. One of the most important traditional methods in SAE is the Fay-Herriot (FH) model, proposed by Fay and Herriot (1979). This method is an areal level model that uses the direct estimators as data, instead of the observed values in the sample. By doing this, it can provide directly estimates for the population. However, we think that a critical analysis must be done about the assumptions in this model. One of them is the normality assumption for the direct estimators. In many real applications that assumption is not adequate.

From 2014 onwards, all the sampling units in the LFS are georeferenced, namely the dwellings in which the observation units (i.e. individuals) are interviewed. This new data structure permit using point referenced models (Banerjee et al, 2005). Hence, with such new information as this, the objective now becomes to model the spatial variation of the sampled marks, namely the number of unemployed people in each of the sampled dwellings, using a point level model, and then to extrapolate in space to all geo-referenced dwellings using spatial smoothing. The number of unemployed people in any areal unit A can then be calculated as the sum of the unemployed in all of the dwellings in that areal unit. Therefore, the suggested modelling strategy, based on 14,000 dwellings sampled in each quarterly sample survey, fits a Poisson generalized linear model with a latent spatio-temporal structured random effect for the number of unemployed people observed in these units, and, by spatial smoothing, extends these unemployment figures to all dwellings in the population whose geo-referenced positions are now known.

In addition to the spatial smoothing in space, we intend to do a temporal extrapolation. The temporal extension will be based on 9 sequentially observed quarterly sampling surveys (from the 4th quarter of 2014 to the 4th quarter of 2016).

For the modelling process, we suggest a geostatistical model with a temporally and spatially strucutured random effect. Typically, in this framework, the spatial process is a Gaussian field (GF). Inference on such models is not straightforward due to the dense covariance matrices, problem known in the literature as *big n problem* (Banerjee et al, 2004). Due to computational problems that emerge in this framework, Lindgren *et al* (2011) proposed a more computationally tractable approach based on stochastic partial differential equation (SPDE) models, which permit the transformation of a Gaussian field to a Gaussian markov random field and we follow this method.

2 Data

We use the Portuguese LFS data from the 4th quarter of 2014 to the 4th quarter of 2016 in the mainland territory. In each quarter, the sample has around 35000 observations, distributed in about 14000 dwellings, which are in about 13800 residential buildings. Thus, in the most part of the sampled residential buildings, only one dwelling is selected. Each individual in the sample are questioned about their state in the labour market (employed, unemployed, inactive), sex, age, education level (primary level, secondary level, higher level), etc.

The georeferencing of all residential buildings are available, even for the units outside the sample. Although a residential building can have multiple dwellings, the coordinates information are available only for the buildings themselves. Since there may be more than one dwelling in each residential unit, particularly in areas of high population density, multiple dwellings in the survey have the same spatial location. To avoid an overlap in the locations within the modelling process, the observation units we will consider are the residential buildings. In the following sections, we will denote the average number of unemployed people per dwelling in the residential building at s_j location and quarter t by $y(s_j,t)$ (rounding to the nearest integer). Here, we intend to extrapolate the values observed in the sampled locations to all residential buildings (around 2300000) by spatial smoothing based on the proposed model.

In the modelling process, we use some covariates at residential buildings level for each quarter, namely the mean age and the median of the education level.

3 Point referenced data models for unemployment estimation

We will assume a Poisson distribution for $y(s,t)$, the average number of unemployed people per dwelling observed at residential building with spatial location at s and in quarterly survey t:

$$y(s,t)|\lambda(s,t) \sim \text{Poisson}(\lambda(s,t)) \tag{1}$$

with

$$\log(\lambda(s,t)) = \alpha + \text{offset}(s,t) + \sum_{m=1}^{M} \theta_m z_m(s,t) + W(s,t), \tag{2}$$

where $W(s,t)$ is a latent spatio-temporal process, $\theta = c(\alpha, \{\theta_m, m = 1,...,M\})$ are the model parameters, $\text{offset}(s,t)$ is the offset term described in the data section and $\{z_m(s,t), m = 1,...,M\}$ are the spatio-temporal covariates.

4 Results

The resultant map of the posterior mean of the average number of unemployed people per dwelling at location s and quarter t, $\lambda(s,t)$, is shown in figure 2. We can see that the average number of unemployed people per dwelling is higher in the Porto, Península de Setúbal and Alentejo regions. We also see a slightly decrease of this indicator across time.

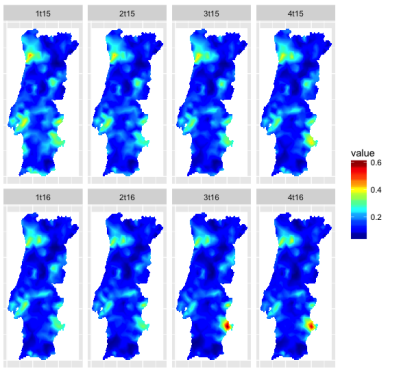


Figure 2: Posterior mean of the average number of unemployed people per dwelling by grid cell

The aggregation of the estimates of the total unemployed by NUTS III regions are shown in figure 3. Here, we can see that there was a decreasing tendency across time during the study period.

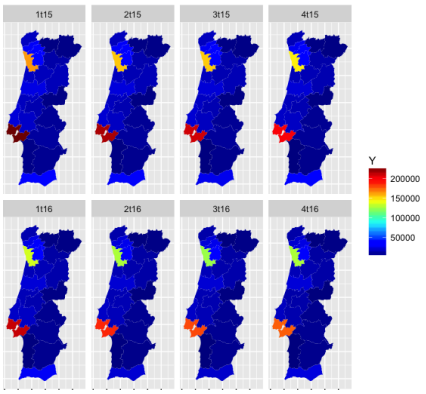


Figure 3: Posterior predictive mean of the total unemployed by NUTS III regions

References

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

[2] Fay, R.E, Herriot,R.A.(1979) Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269-277.

[3] Lindgren, F., Rue, H., Lindstrom, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the SPDE approach (with discussion). *Journal of Royal Statistical Society Series B*, **73**, 423-498.

[4] Rao, J.N.K., Molina, I. (2015) *Small Area Estimation*, Second edition. New York: Wiley.

Thin-plate splines for cloud filling in satellite imagery

Unai Pérez-Goya[1,*], Ana F Militino[1], M Dolores Ugarte[1] and Marc Genton [2]

[1] *Department of Statistics and Operations Research, Public University of Navarre (Spain).*
E-mails: unai.perez@unavarra.es; militino@unavarra.es, lola@unavarra.es
[2] *Computer, Electrical and Mathematical Sciences and Engineering (CEMSE), Kingdom of Saudi Arabia.*
E-mail: marc.genton@kaust.edu.sa

Abstract. *Time series of satellite imagery is nowadays essential for monitoring the evolution of land use, land cover, vegetation trends, and climatological or phenological changes. However, but some of these images could be useless because of the presence of clouds. In this paper we propose filling these clouds through a thin-plate spline (Tps) model that accommodates spatio-temporal dependence among images. The performance of the method is illustrated with a simulation study where the Tps procedure is compared with other alternatives. The scenario of the simulation study consists in introducing at random seven sizes of clouds in three time series of composite MODIS Terra and Aqua images of Navarre (Spain) between 2011 and 2013. The remote sensing data are the normalized difference vegetation index (NDVI) and the land surface temperature (LST) day and night. The results show that the thin-plate spline model outperforms Timesat, Hants and Gapfill in small, moderate, and big clouds in LST day and night and it is equally competitive with NDVI.*

Keywords. *kriging; cloud-filling; satellite imagery; geostatistics*

---

1 Introduction

Numerous cloud-filling and smoothing techniques have been developed in the past few years for filling clouds in satellite imagery [2, 10]. For example, Timesat [3, 4] or Hants [5] are very popular because of its good performance and free access, and Gapfill [6] is a very recent and promising alternative. These methods are based on harmonic analysis, polynomial functions with different filters, and quantile regression respectively, and with different formats all of them borrow the similarity among close images to fill the clouds. In this work the Normalized Difference Vegetation Index (NDVI) and the Land Surface Temperature (LST) day and night are used to illustrate the advantages of modelling with thin-plate splines in a pre-defined neighbourhood of the target image, that includes previous and subsequent images across the years in a similar way as it is in Gapfill.

2 Data

In this paper we use composite images [1] of the Navarre region from the 2011-2013 time period. Navarre is a region of approximately $10,000\ km^2$ located in the north of Spain. Elevations vary between 200 and

2,500 meters in the highest zone of the Pyrenees, located in Northeastern Navarre. Hence, a high presence of clouds is expected in the north of the region, particularly in Winter [7]. The LST day and night composite images are provided from the Moderate Resolution Imaging Spectroradiometer (MODIS). Each one of the LST day and LST night remote sensing data require 138 tiles for enclosing Navarre. The dataset correspond to 46 scenes with a temporal resolution of 8 days every year. NDVI images of MODIS Aqua and Terra are composed every 16 days, so that only 23 images are available every year. Therefore, to achieve 2 composite NDVI images per month every year, we have retrieved 23 images from Aqua and one image from Terra. In total, 72 scenes with a 16-day temporal resolution have been captured across 3 years of study. The spatial resolution of each tile is equal to 156×145 (22620) pixels of approximately 1 km^2 but 11691 pixels are needed to enclose Navarre region.

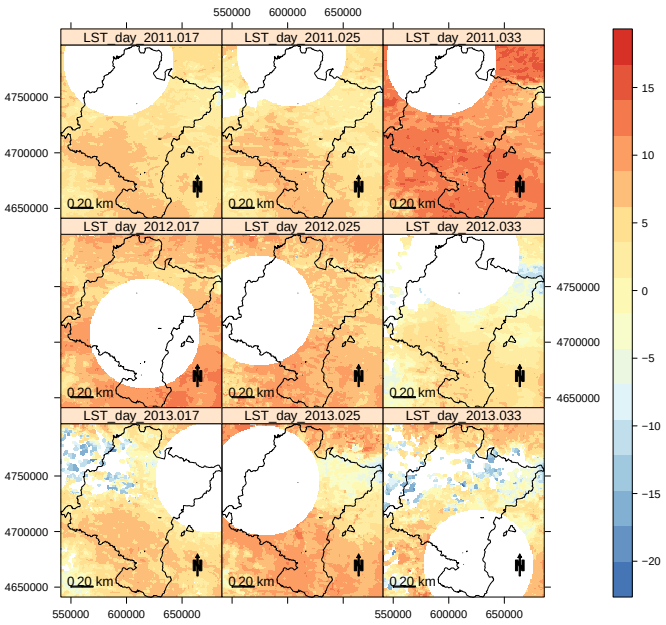


Figure 1: Neighborhood example of the LST_day 2011_025 image involved in the thin-plate model and gapfill method, where different random gaps of size I have been introduced in every image

3 The thin-plate spline model

The thin-plate spline model (Tps) [8] is applied to the mean residuals derived from a neighbourhood of the target image. For defining the neighbour, let us start with an LST day target image. In this case, $G = 46$ images are available every year from ($r = 2011, \ldots, 2013$), which should be arranged into a $3 \times G = 3 \times 46$ matrix, where the rows of the matrix correspond to different years. All the images in the same column correspond to the same time period but different years. They share a neighbour composed of this column and the previous and subsequent columns of images; therefore, the neighbour of every target image consists of 9 images. See Figure 1 as a particular example of the LST day image of the 25th Julian day of 2012 and its neighbourhood. In the first time period of 2011 and in the last time period of 2013, previous and subsequent images can be used, yet they do not correspond to the years under study. The second step of this procedure is to compute the mean image (z_{0g}) out of those 9 images and obtain

the corresponding residuals for the target image ($w_{\mathbf{s}rg}$) from $(s = s_1, \ldots, s_n)$, where

$$w_{s_i rg} = z_{s_i rg} - z_{0g}. \tag{1}$$

Next, a thin-plate spline model (Tps) is applied to a 5-times lower resolution of these residuals, yet the resolution level could change depending on the computing capacity. The lower resolution is obtained through a mean aggregation. The thin-plate spline model is expressed as a non-parametric function of the coordinates plus the normal error term, so that

$$\mathbf{w}_{st_j} = f(\mathbf{x}_s, \mathbf{y}_s) + \boldsymbol{\varepsilon}_{st_j}, \tag{2}$$

where $\mathbf{w}_{st_j} = (w_{s_1 t_j}, \ldots, w_{s_n t_j})'$ are the n observed pixels (or remote sensing data) of any image captured in time t_j, and $\mathbf{x}_s, \mathbf{y}_s$ are the planar coordinates depending on the locations $s = (s_1, \ldots, s_n)$. The error $\boldsymbol{\varepsilon}_{st}$ is assumed to be normally distributed $\boldsymbol{\varepsilon}_{st} \sim N(\mathbf{0}, \Sigma(d))$, where $\Sigma(d)$ is the covariance matrix depending on the Euclidean distance d. The spline is obtained as a weighted average of the observed data because the optimal estimate of $f(\mathbf{x}_s, \mathbf{y}_s)$ turns out to be linear in the observations. Finally, the predictions $\hat{z}_{srg} = z_{srg} + z_{0g}$ are computed in the original resolution. The thin-plate spline model is fitted in the R statistical software using the fields package [9].

4 Results

The simulation study consists in running the 6 methods: 3 versions of Timesat, Hants, Gapfill, and Tps with 7 sizes of artificial clouds (C, D, E, F, G, H, I) that produce missing data inside a circle. Timesat needs to be run only once for every remote sensing data and for every size of the artificial clouds in the three years. Therefore, the introduction of artificial clouds is done once for the whole period, size cloud and derived variable, yet every image can have different random clouds. Hants needs to introduce three series of random clouds every year, every size of cloud, and every one of the derived variables, because the method is run every year separately, otherwise it over-smooths the images. Gapfill requires pre-defined neighbourhood of 9 images for smoothing 3 images and then, for every running, cloud size, and derived variable, 9 random clouds are introduced. The performance of the methods is evaluated with the square root of the mean squared prediction error calculated as the square root of the mean square differences between observed and filled data for every derived variable in every period and cloud size. The study shows that overall Tps outperforms the other alternatives, except in big sizes of artificial clouds, where Gapfill can be more competitive that Tps. However, Gapfill is slower than Tps and could crash when similar pixels are found in the pre-defined neighbourhoods and ranks are not well balanced.

References

[1] Holben, B. N. (1986). Characteristics of maximum-value composite images from temporal AVHRR data. International journal of remote sensing, 7(11), 1417-1434.

[2] Militino, A.F.; Ugarte, M.D.; Pérez-Goya, U. Stochastic Spatio-Temporal Models for Analysing NDVI Distribution of GIMMS NDVI3g Images. Remote Sensing 2017, 9, 76.

[3] Eklundh, L., & Jönsson, P. (2012). TIMESAT 3.2 with parallel processing-Software Manual. Lund University.

[4] Jönsson, P.; Eklundh, L. TIMESAT a program for analyzing time-series of satellite sensor data. Computers & Geosciences 2004, 30, 833–845.

[5] Verhoef, W., Menenti, M., & Azzali, S. (1996). Cover A colour composite of NOAA-AVHRR-NDVI based on time series analysis (1981-1992). International Journal of Remote Sensing, 17(2), 231-235.

[6] Gerber, F., de Jong, R., Schaepman, M. E., Schaepman-Strub, G., & Furrer, R. (2018). Predicting Missing Values in Spatio-Temporal Remote Sensing Data. IEEE Transactions on Geoscience and Remote Sensing.

[7] Militino, A.; Ugarte, M.; Goicoa, T.; Genton, M. (2015). Interpolation of daily rainfall using spatiotemporal models and clustering. International Journal of Climatology 35, 1453–1464.

[8] Wahba, G. Spline models for observational data; CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), 1990.

[9] *Fields: Tools for Spatial Data*. R Package Version 9.0; R Foundation for Statistical Computing: Vienna, Austria, 2015.

[10] Yin, G., McCabe, M. F., Mariethoz, G. and Sun, Y. (2017), Comparison of gap-filling methods for Landsat 7 ETM+ SLC-off imagery. International Journal of Remote Sensing, 38(23), 6653-6679.

Estimation of spatial autoregressive conditional heteroscedasticity models

P. Otto[1,*], W. Schmid[1] and R. Garthoff[2]

[1] *European University Viadrina, Frankfurt (Oder); potto@europa-uni.de*
[2] *Statistisches Landesamt des Freistaates Sachsen, Kamenz*
[*] *Corresponding author*

Abstract. *Otto, Schmid, and Garthoff (2016) introduce a new spatial model that incorporates heteroscedastic variance depending on neighboring locations. The proposed process is regarded as the spatial equivalent to the temporal autoregressive conditional heteroscedasticity (ARCH) model. In contrast to the temporal ARCH model, in which the distribution is known given the full information set of the prior periods, the distribution is not straightforward in spatial and spatiotemporal settings. However, it is possible to estimate the parameters of the model using the maximum-likelihood approach. Moreover, we introduce an exponential spatial ARCH models and propose a maximum-likelihood estimator for this kind of spatial ARCH model. In the talk, I focus on the estimation from a computational and practical point of view. From this perspective, the log-likelihood function is usually sufficient to get accurate parameter estimates by using any non-linear, numerical optimization function. To compute the likelihood for a certain set of parameters, the determinant of the Jacobian matrix must be computed, which often requires large computationally capacities, especially for large data sets. In particular, I show the implementation of the estimation approach in the R-package* `spGARCH`*. Eventually, the function for estimation is demonstrated by an illustrative example.*

Keywords. *Computational statistics; Disease mapping; Spatial and spatio-temporal covariance modelling.*

---

1 Model Definition

The class of spatial ARCH models have been introduced by [13]. In particular, we consider a univariate stochastic process $\{Y(\boldsymbol{s}) \in R : \boldsymbol{s} \in D_{\boldsymbol{s}}\}$ having a spatial autoregressive structure in the conditional variance. The process is defined in a multidimensional space $D_{\boldsymbol{s}}$, which could be a subset of the q-dimensional real numbers R^q or of the q-dimensional integers Z^q. For the first case, it is important that the subset contains q-dimensional rectangle of positive volume (cf. [4]). In the latter case, the process is called spatial lattice process. Moreover, this definition is suitable to model spatiotemporal data, as one might assume that $D_{\boldsymbol{s}}$ is a product $R^k \times Z^l$ with $k+l=d$.

Let $\boldsymbol{s}_1,\ldots,\boldsymbol{s}_n$ denote all locations, and let $\boldsymbol{Y}$ be the vector of observations $(Y(\boldsymbol{s}_i))_{i=1,\ldots,n}$, which is given by

$$\boldsymbol{Y} = \operatorname{diag}(\boldsymbol{h})^{1/2}\boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{\varepsilon} = (\varepsilon(\boldsymbol{s}_1),\ldots,\varepsilon(\boldsymbol{s}_n))'$ is assumed to be an independent and identically distributed random error

with $E(\varepsilon) = \mathbf{0}$ and $Cov(\varepsilon) = \mathbf{I}$. The identity matrix is denoted by $\mathbf{I}$. The representation is analogous to the ARCH time-series model of [6]. We now distinguish between several spatial ARCH-type models via the definition of $\boldsymbol{h}$.

1.1 Special ARCH-type Models

First, we define this vector $\boldsymbol{h}$ analogous to the definition in [13]. For this model, the vector $\boldsymbol{h}_O$ is given by

$$\boldsymbol{h}_O = \alpha\mathbf{1} + \rho\mathbf{W}\text{diag}(\boldsymbol{Y})\boldsymbol{Y}\,, \tag{2}$$

where $\text{diag}(\boldsymbol{a})$ is a diagonal matrix with the entries of $\boldsymbol{a}$ on the diagonal.

It is important to assume that the spatial weighting matrix is non-stochastic, positive matrix, which has zeros on the main diagonal to assure that a location is not influenced by itself (cf. [5, 4]). The support of the distribution of the random errors ε must be compact under certain conditions. Due to the complex dependence implied by the weighting matrix $\mathbf{W}$, $\boldsymbol{h}_O$ is not necessarily positive; thus, $\text{diag}(\boldsymbol{h})^{1/2}$ might not have a solution in the real numbers.

There are two cases where the support of the errors does not need to be constraint. If $\rho = 0$, the process coincides with a spatial white noise process. Moreover, all entries of $\boldsymbol{h}$ are non-negative, if $\mathbf{W}$ is similar to a strictly triangular matrix. This is the case, if $\mathbf{W}$ is nilpotent. This case covers the classical time-series ARCH(p) models introduced by [6] as well as so-called oriented spARCH processes. For these processes, the spatial dependence has a certain direction, e.g. observations are only influenced by observations in a southward direction, or by observation which are closer to an arbitrary center. This setting also covers recent time-series GARCH models incorporating spatial information (e.g. [2, 3]).

Secondly, we consider an exponential spatial ARCH process (E-spARCH). In this setting, we define the logarithm of $\boldsymbol{h}_E$ as

$$\ln \boldsymbol{h}_E = \alpha\mathbf{1} + \rho\mathbf{W}g_b(\varepsilon)\,, \tag{3}$$

with a function $g_b : R^n \to R^n$. Like [12], we assume that

$$g_b(\varepsilon) = (\ln|\varepsilon(\boldsymbol{s}_1)|^b, \ldots, \ln|\varepsilon(\boldsymbol{s}_n)|^b)'$$

for positive values of b. At location $\boldsymbol{s}_i$, the value of $h_E(\boldsymbol{s}_i)$ is then given by

$$\ln h_E(\boldsymbol{s}_i) = \alpha + \sum_{v=1}^{n} \rho b w_{iv} \ln|\varepsilon(\boldsymbol{s}_v)| \text{ for } i = 1, \ldots, n\,. \tag{4}$$

For this definition of g_b, one could rewrite $\ln \boldsymbol{h}$ as

$$\ln \boldsymbol{h}_E = \mathbf{S}\,(\alpha\mathbf{1} + \rho b\mathbf{W}\ln|\boldsymbol{Y}|) \tag{5}$$

with

$$\mathbf{S} = (s_{ij})_{i,j=1,\ldots,n} = \left(\mathbf{I} + \frac{1}{2}\rho b\mathbf{W}\right)^{-1}.$$

Thirdly, we focus on a mixed model. In particular, a spatial autoregressive (SAR) model with spatial ARCH residuals is considered. To define the SAR process, we must introduce a further matrix $\mathbf{B}$ of spatial weights. This matrix $\mathbf{B}$ could differ from the aforementioned weighting matrix $\mathbf{W}$. However, it $\mathbf{B}$

is also assumed to be non-stochastic and nonnegative with zeros on the main diagonal. Furthermore, let λ denote the SAR coefficient, and let $\mathbf{X}\boldsymbol{\beta}$ be a regressive term. The model is then defined as follows

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \lambda\mathbf{B}\boldsymbol{Y} + \boldsymbol{\xi}\ \text{,i.e.}\quad \boldsymbol{Y} = (\mathbf{I} - \lambda\mathbf{B})^{-1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi})\,. \tag{6}$$

The vector of disturbances $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ follows a spARCH according to the suggested model in (1).

Compared to an SAR process with autoregressive or moving average residuals (SARAR/SARMA; cf. [11, 7, 9]), the proposed SAR model with spARCH errors (SARspARCH) has a great advantage. The first approach also makes it possible to model spatial heteroscedasticity with an autoregressive error term. However, the autoregressive structure of the errors also affects the spatial autocorrelation of the process. For the proposed SARspARCH model, it is possible to model the spatial dependence of the process solely by the weight matrix of the autoregressive part and the heteroscedasticity by the weight matrix of the spARCH error term. From a practical perspective, this approach has the great advantage that a spARCH error term can easily be incorporated if a suitable weighting scheme for the spatial mean process is found, but the residuals remain heteroscedastic. The spatial mean process will not be affected by the spARCH errors. Generally, the choice of the weighting matrices $\mathbf{B}$ and $\mathbf{W}$ depends on the specific problem. Accordingly, if the spatial mean and the variance process are related to each other, one would expect a similar structure for both weight matrices, e.g., spatial contiguity matrices. In contrast, one might choose completely different weighting schemes if the spatial mean process and the spatial variance are not related to each other.

2 Parameter Estimation

The parameters of a spatial ARCH process can be estimated via by the maximum-likelihood approach. To obtain the joint density for $\boldsymbol{Y} = f(\varepsilon)$, the Jacobian matrix of f^{-1} at the observed values $\boldsymbol{y}$ has to be computed (e.g., [1]). If $f_{\boldsymbol{\varepsilon}}$ is the distribution of the independent random errors, the joint density $f_{\boldsymbol{Y}}$ of $\boldsymbol{Y}$ is given by

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= f_{(Y(\boldsymbol{s}_1),\ldots,Y(\boldsymbol{s}_n))}(y_1,\ldots,y_n) \\ &= f_{\boldsymbol{\varepsilon}}\left(\frac{y_1}{\sqrt{h_1}},\ldots,\frac{y_n}{\sqrt{h_n}}\right)\left|\det\left(\left(\frac{\partial y_j/\sqrt{h_j}}{\partial y_i}\right)_{i,j=1,\ldots,n}\right)\right|. \end{aligned} \tag{7}$$

Consequently, the parameters can be estimated by the maximum-likelihood approach. The parameter estimates are obtained from the maximization of the log likelihood, i.e.,

$$(\hat{\alpha},\hat{\rho}) = \underset{\alpha>0,\rho\geq 0}{\arg\max}\ \ln\left|\det\left(\left(\frac{\partial y_j/\sqrt{h_j}}{\partial y_i}\right)_{i,j=1,\ldots,n}\right)\right| + \sum_{i=1}^{n}\ln f_{\varepsilon}(y_i)\,.$$

The Jacobian matrix, of course, depends on the definition of $\boldsymbol{h}$.

In the `spGARCH` package, we implemented the iterative maximization algorithm proposed by [14], which is implemented in the `R`-package `Rsolnp` (see [8]). However, note that the first determinant of the log determinant of the Jacobian also depends on the parameters, such that it needs to be computed in each iteration (see also Theorem 13.7.3 of [10] for the computation of a determinant of the sum of a diagonal matrix and an arbitrary matrix), but $\mathbf{W}$, and therefore also $\mathbf{S}\circ\mathbf{W}$, are usually sparse.

3 Implementation in the R-package `spGARCH`

The `R`-package `spGARCH` provides several basic function for the analysis of spatial data showing spatial conditional heteroscedasticity. In particular, the process can be simulated for arbitrary weighting matrices according to the definitions given above. Moreover, we implement a function for the estimation of the model parameters by the maximum-likelihood approach. To generate a user-friendly output, the object generated by the estimation function can easily be summarized by the generic function `summary`. We also provide all common generic methods like `plot`, `print`, `logLik`, etc. To maximize the computational efficiency, actual version of the packages contains compiled `C++` code.

References

[1] Peter J Bickel and Kjell A Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*, volume 117. CRC Press, 2015.

[2] Svetlana Borovkova and Rik Lopuhaa. Spatial GARCH: A Spatial Approach to Multivariate Volatility Modeling. *Available at SSRN 2176781*, 2012.

[3] Massimiliano Caporin and Paolo Paruolo. GARCH Models with Spatial Structure. *SIS Statistica*, pages 447–450, 2006.

[4] Noel Cressie and Christopher K Wikle. *Statistics for Spatio-Temporal Data*. Wiley, 2011.

[5] J Paul Elhorst. Applied Spatial Econometrics: Raising the Bar. *Spatial Economic Analysis*, 5(1):9–28, 2010.

[6] Robert F Engle. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007, 1982.

[7] Bernard Fingleton. A Generalized Method of Moments Estimator for a Spatial Panel Model with an Endogenous Spatial Lag and Spatial Moving Average Errors. *Spatial Economic Analysis*, 3(1):27–44, 2008.

[8] Alexios Ghalanos and Stefan Theussl. *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*, 2012. R package version 1.14.

[9] R. P. Haining. The Moving Average Model for Spatial Interaction. *Transactions of the Institute of British Geographers*, 3(2):202–225, 1978.

[10] David A Harville. *Matrix Algebra from a Statistician's Perspective*, volume 1. Springer, 2008.

[11] Harry H Kelejian and Ingmar R Prucha. Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances. *Journal of Econometrics*, 157(1):53–67, 2010.

[12] Daniel B Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, pages 347–370, 1991.

[13] Philipp Otto, Wolfgang Schmid, and Robert Garthoff. Generalized spatial and spatiotemporal autoregressive conditional heteroscedasticity. Technical report, Discussion Paper Series, European University Viadrina, Frankfurt (Oder), 2016.

[14] Yinyu Ye. *Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming*. PhD thesis, Department of ESS, Stanford University, 1988.

A multilevel hidden Markov model for space-time cylindrical data

Francesco Lagona[1] and Monia Ranalli[1,*]

[1] *Department of Political Sciences, Roma Tre University (Italy); francesco.lagona@uniroma3.it, monia.ranalli@uniroma3.it*
[*] *Corresponding author*

Abstract. *Motivated by segmentation issues in marine studies, a novel hidden Markov model is proposed for the analysis of cylindrical space-time series, that is, bivariate space-time series of intensities and angles. The model is a multilevel mixture of cylindrical densities, where the parameters of the mixture vary at the spatial level according to a latent Markov random field, while the parameters of the hidden Markov random field evolve at the temporal level according to the states of a hidden Markov chain. It segments the data within a finite number of latent classes that represent the conditional distributions of the data under environmental conditions that vary across space and time, simultaneously accounting for unobserved heterogeneity and space-time autocorrelation. It parsimoniously accommodates specific features of environmental cylindrical data, such as circular-linear correlation, multimodality and skewness. Due to the numerical intractability of the likelihood function, parameters are estimated by a computationally efficient EM algorithm based on the maximization of a weighted composite likelihood. The effectiveness of the proposal is tested in a case study that involves speeds and directions of marine currents in the Gulf of Naples, where the model was capable to cluster cylindrical data according to a finite number of intuitively appealing latent classes.*

Keywords. *Cylindrical data; hidden Markov model; EM algorithm; Composite likelihood*

---

1 Introduction

A detailed knowledge of coastal currents is crucial for a valid integrated coastal zone management. Among the different available ocean observing technologies, high-frequency radars (HFRs) have unique characteristics, that make them play a key role in coastal observatories. HFR data can be conveniently described as space-time bivariate arrays of angles and intensities that respectively indicate the directions and the speeds of sea currents across space and over time. Data with a mixed circular-linear support are often referred to as *cylindrical* data (Abe and Ley, 2017), because the pair of an angle and an intensity can be represented as a point on a cylinder.

The statistical analysis of cylindrical space-time series is complicated by the unconventional topology of the cylinder and by the difficulties in modeling the cross-correlations between angular and linear measurements across space and over time. Additional complications arise from the skewness and the multimodality of the marginal distributions of the data. As a result, specific methods for the analysis of space-time cylindrical data have been relatively unexplored. Proposals in this context are limited to geostastical models, where cylindrical data are assumed conditionally independent given a latent process

that varies continuously across space and time (Wang et al., 2015). Geostatistical models give good results in open sea areas, where waves and currents can move freely without obstacles. Sea motion in coastal areas provides, however, a different setting. Coastal currents are shaped and constrained by the orography of the site. As a result, coastal circulation is much more irregular than ocean-type patterns and it is inaccurately represented by traditional geostatistical models, which do not incorporate orographic information. The development of a physical model that well represents sea motion in coastal areas can be a formidable task if the orography of the site is irregular. A more practical approach relies on decomposing an observed circulation pattern into a small number of local regimes whose interpretation is easier than the global pattern.

To accomplish this goal, we propose a model that segments coastal data according to finitely many latent classes that vary across space and time and are associated with the distribution of the data under specific, space-time varying, environmental conditions. Specifically, we assume that the joint distribution of the data is well approximated by a multi-level mixture of cylindrical densities. At each time, the parameters of the mixture vary according to a latent Markov field, whose parameters evolve over time according to a latent Markov chain. The idea of using hidden Markov models to segment cylindrical data is not completely novel. Lagona et al. (2015) propose a hidden Markov model for segmenting cylindrical time series, while Ranalli et al. (2018) propose a hidden Markov field to segment spatial cylindrical data. Our proposal integrates these specifications in a space-time setting.

A potential disadvantage of the model is the intractability of the likelihood function. We address estimation issues by relying on composite likelihood (CL) methods (Varin et al., 2011; Lindsay, 1988). This estimation strategy, on one hand, provides feasible and fast estimation methods. On the other hand, some dependence among observations is lost, resulting in a loss of statistical efficiency. However, consistency of the CL estimators still holds under regularity conditions (Molenberghs and Verbeke, 2005). Under these conditions, furthermore, CLEs are asymptotically normal with covariance matrix given by the inverse of a sandwich matrix, known as Godambe information (Godambe, 1960) rather than the usual Fisher information matrix for maximum likelihood estimators (MLEs). CL methods have been successfully applied in spatial and space-time settings (Ranalli et al., 2018; Okabayashi et al., 2011; Eidsvik et al., 2014).

2 A cylindrical space-time hidden Markov model

The data that motivated this work are in the form of an $n \times T$ array of cylindrical data, say $(\mathbf{z}_{it}, i = 1 \dots n, t = 1 \dots T)$, where $\mathbf{z}_{it} = (x_{it}, y_{it})$ is a pair of an angle $x_{it} \in [0, 2\pi)$ and an intensity $y_{it} \in [0, +\infty)$, observed at time t and in the spatial site i. We assume that the temporal evolution of these data is driven by a multinomial process in discrete time $\boldsymbol{\xi} = (\boldsymbol{\xi}_t, t = 1 \dots T)$, where $\boldsymbol{\xi}_t = (\xi_{t1} \dots \xi_{tK})$ is a multinomial random variable with K classes. We specifically assume that such process is distributed as a Markov chain, whose distribution, say $p(\boldsymbol{\xi}; \boldsymbol{\pi})$, is known up to a vector of parameters $\boldsymbol{\pi}$ that includes the initial probabilities and the transition probabilities of the chain. Conditionally on the value assumed each time by the Markov chain, the spatial distribution of the data at time t depends on a multinomial process in discrete space $\mathbf{u}_t = (\mathbf{u}_{it}, i = 1 \dots n)$, where $\mathbf{u}_{it} = (u_{it1}, \dots u_{itG})$ is a multinomial variable with G classes. We assume that such spatial process is distributed as a G-parameter Potts model, whose parameters depend on the value taken at time t by the latent Markov chain $p(\boldsymbol{\xi}; \boldsymbol{\pi})$. This model depends on $G-1$ sufficient statistics

$$n_g(\boldsymbol{u}_t) = \sum_{i=1}^{n} u_{itg}, \quad g = 1 \dots G-1,$$

that indicate the frequencies of each latent class across the study area, and one sufficient statistic

$$n(\boldsymbol{u}_t) = \sum_{i=1}^{n} \sum_{j>i:\, j\in N(i)} \sum_{g=1}^{G_{t-1}} u_{itg} u_{jtg},$$

which indicates the frequency of neighboring sites which share the same class (for each i, $N(i)$ indicates the sets of neighboring sites of i). Precisely, we assume that the joint distribution of a sample $\mathbf{u}_t$, conditionally on $\boldsymbol{\xi}_t$, is known up to an array of class-specific parameters $\boldsymbol{\alpha} = (\alpha_{gk}, g = 1 \dots G-1, k = 1 \dots K)$ and a vector of auto-correlation parameters $\boldsymbol{\rho} = (\rho_1 \dots \rho_k)$, and given by

$$p(\boldsymbol{u}_t \mid \boldsymbol{\xi}_t; \boldsymbol{\alpha}, \boldsymbol{\rho}) = \frac{\exp\left(\sum_{g=1}^{G_{t-1}} n_g(\boldsymbol{u}_t)\alpha_{gt} + n(\boldsymbol{u}_t)\rho_t\right)}{W(\boldsymbol{\alpha}, \boldsymbol{\rho})}, \tag{1}$$

where

$$\alpha_{gt} = \sum_{k=1}^{K} \xi_{tk}\alpha_{gk}$$

and

$$\rho_t = \sum_{k=1}^{K} \xi_{tk}\rho_k.$$

Our proposal is completed by assuming that, conditionally on the values taken by the Markov chain and the Potts model, the observed cylindrical data are independently distributed according to cylindrical densities, known up to a vector of parameters that depends on the latent class taken by the latent Markov random field at time t in site i. Precisely, we assume that

$$f(\mathbf{z} \mid \boldsymbol{\xi}, \mathbf{u}) = \prod_{i=1}^{n}\prod_{t=1}^{T} f(\mathbf{z}_{it}; \boldsymbol{\theta}_{itg}),$$

where

$$\boldsymbol{\theta}_{itg} = \sum_{g=1}^{G} u_{itg}\boldsymbol{\theta}_g,$$

and $\boldsymbol{\theta}_g$ is the gth entry of a vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1 \dots \boldsymbol{\theta}_G)$. Under this setting, we follow Abe and Ley (2017) and exploit the following parametric cylindrical distribution, namely

$$f(\mathbf{z}; \boldsymbol{\theta}) = \frac{\alpha\beta^{\alpha}}{2\pi\cosh(\kappa)}(1 + \lambda\sin(x-\mu))y^{\alpha-1}\exp(-(\beta y)^{\alpha}(1 - \tanh(\kappa)\cos(x-\mu))), \tag{2}$$

known up to five parameters $\boldsymbol{\theta} = (\alpha, \beta, \kappa, \lambda, \mu)$, where $\alpha > 0$ is a shape parameter, $\beta > 0$ is a scale parameter, $\mu \in [0, 2\pi)$ is a circular location parameter, $\kappa > 0$ is a circular concentration parameter, while $\lambda \in [-1, 1]$ is a circular skewness parameter.

The joint distribution of the observed and the latent variables is therefore given by

$$f(\mathbf{z}, \mathbf{u}, \boldsymbol{\xi}; \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\rho}) = f(\mathbf{z} \mid \boldsymbol{u}; \boldsymbol{\theta}) p(\mathbf{u}; \boldsymbol{\rho}, \boldsymbol{\alpha}) p(\boldsymbol{\xi}; \boldsymbol{\pi}). \tag{3}$$

By integrating this distribution with respect to the unobserved variables, we obtain the likelihood function of the unknown parameters

$$L(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\rho}; \mathbf{z}) = \sum_{\boldsymbol{\xi}} \sum_{\boldsymbol{u}} f(\mathbf{z}, \boldsymbol{u}, \boldsymbol{\xi}; \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\rho}). \tag{4}$$

The maximization of the corresponding complete log-likelihood through an EM algorithm is unfeasible. As a result, we propose to estimate the parameters by maximizing a surrogate function, namely a composite log-likelihood function. Our proposal is based on the specification of a cover $\mathbb{A}$ of the set $S = \{1 \dots n\}$ of the observation sites, i.e. a family of (not necessarily disjoint) subsets $A \subseteq S$ such that $\bigcup_{A \in \mathbb{A}} = S$. For each subset A, we respectively define $\mathbf{z}_A = (\mathbf{z}_{it}, i \in A, t = 1 \dots T)$, $\mathbf{u}_A = (\mathbf{u}_{it}, i \in A, t = 1 \dots T)$, and

$$L^A(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\rho}; \mathbf{z}_A) = \sum_{\boldsymbol{\xi}} \sum_{\boldsymbol{u}_A} f(\mathbf{z}_A, \boldsymbol{u}_A, \boldsymbol{\xi}; \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\rho}) \tag{5}$$

as the contribution of the data in A to the composite likelihood (CL), where CL=$\prod_{A \in \mathbb{A}} L^A$. This composite likelihood function involves summations over all the possible values that $\boldsymbol{u}_A$ can take. As a result, the numerical tractability of these steps dramatically decreases with the cardinality of the largest subset of the cover $\mathbb{A}$. On the one side, this would suggest to choose a cover with many small subsets. On the other side, a cover that includes a few large subsets is expected to provide a CL function that is a better approximation of the likelihood function. Because summations over $\boldsymbol{u}_A$ become cumbersome for $|A| \geq 3$, a natural strategy is a cover that includes subsets with 2 elements. When $\mathbb{A}$ include all the subsets of two elements, then composite likelihood reduces to the pairwise likelihood function (Varin et al., 2011). In a spatial setting, a pairwise likelihood can be further simplified by discarding all the pairs (i, j) that are not in the neighborhood structure $N(i), i = 1 \dots n$. This choice provides a computationally efficient EM algorithm, without sacrificing the good distributional properties that are expected by a CL estimator.

References

Abe, T. and C. Ley (2017). A tractable, parsimonious and flexible model for cylindrical data, with applications. *Econometrics and Statistics 4*, 91 – 104.

Eidsvik, J., B. A. Shaby, B. J. Reich, M. Wheeler, and J. Niemi (2014). Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics 23*(2), 295–315.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics 31*(4), pp. 1208–1211.

Lagona, F., M. Picone, and A. Maruotti (2015). A hidden markov model for the analysis of cylindrical time series. *Environmetrics 26*, 534–544.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics 80*, 221–239.

Molenberghs, G. and G. Verbeke (2005). *Models for discrete longitudinal data*. Springer Series in Statistics Series. Springer Science+Business Media, Incorporated New York.

Okabayashi, S., L. Johnson, and C. Geyer (2011, 1). Extending pseudo-likelihood for potts models. *Statistica Sinica 21*(1), 331–347.

Ranalli, M., F. Lagona, M. Picone, and E. Zambianchi (2018). Segmentation of sea current fields by cylindrical hidden Markov models: a composite likelihood approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 67*(3), 575–598.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica 21*(1), 1–41.

Wang, F., A. Gelfand, and G. Jona-Lasinio (2015). Joint spatio-temporal analysis of a linear and a directional variable: space-time modeling of wave heights and wave directions in the adriatic sea. stat. sinica. *Statistica Sinica 25*, 25–39.

Simulation of isotropic Gaussian random Fields on Spheres

C. Lantuéjoul

MinesParisTech, 35 rue Saint-Honoré, 77305 Fontainebleau, France; christian.lantuejoul@mines-paristech.fr

---

Abstract. *In several domains of the Geosciences (climatology, cosmology, geodesy or paleomagnetism), data are supported by spheres. As they often exhibit spatial strutures, it may be interesting to examine them using a geostatistical approach. At first some background is provided on several geometric and stochasic features of the sphere that make it so different from Euclidean spaces (spherical harmonics, Schoenberg representation of covariance functions). Then an algorithm is proposed to perform continuous simulations of isotropic Gaussian random fields on the sphere. This algorithm requires knowledge of the spectral measure of the covariance function. Besides a few examples, particular attention is paid to the spectral measures of the Yadrenko class of covariance functions.*

Keywords. *Spherical harmonics; Spectral measures; Yadrenko covariances; Chentsov construction.*

---

1 Background: Geometric aspects

In this presentation, the workspace is the unit sphere $\mathbb{S}^2 = \{x \in \mathbb{R}^3 : |x| = 1\}$ centered at the origin, say o. Each point $x \in \mathbb{S}^2$ can be specified by its *polar coordinates*, namely its *longitude* $0 \leq \phi < 2\pi$ and its *colatitude* $0 \leq \theta \leq \pi$. The *geodesic distance* between two points $x, y \in \mathbb{S}^2$ is the angle $\alpha(x,y)$ separating x and y when both points are seen from o. Explicitly $\alpha(x,y) = \arccos x \cdot y$, where $x \cdot y$ denotes the scalar product between $\vec{ox}$ and $\vec{oy}$ in $\mathbb{R}^3$. On the metric space $(\mathbb{S}^2, \alpha)$ the *invariant measure* $d\sigma(x) = \sin\theta\, d\theta\, d\phi$ can be defined. Its integral over $\mathbb{S}^2$ is equal to 4π.

Spherical harmonics are complex-valued functions that act on $\mathbb{S}^2$ exactly as complex exponentials on the circle. They are defined by

$$Y_{n,k}(x) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-k)!}{(n+k)!}}\, P_{n,k}(\cos\theta)\, e^{ik\phi} \qquad x = (\theta,\phi) \in \mathbb{S}^2,$$

where

$$P_{n,k}(t) = \frac{(-1)^{n+k}}{2^n n!}(1-t^2)^{k/2} D^{n+k}(1-t^2)^n \qquad -1 \leq t \leq +1$$

is the associated Legendre polynomial. In particular, $P_{n,0}$ is a Legendre polynomial (noted P_n). At first, it may look strange to index a countable family of functions using a *degree* n and an *order* k. In fact, n is non-negative and k varies from $-n$ to $+n$, so that $Y_{n,-k} = (-1)^k \bar{Y}_{n,k}$. Fig. 1 shows typical examples of spherical harmonics.

Spherical harmonics satisfy two properties [2] that are essential in this presentation:

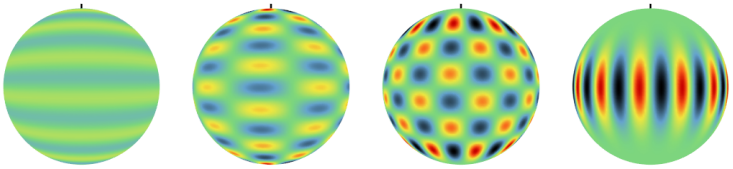


Figure 1: Spherical harmonics of degree 15 and orders $0, 5, 10$ et 15 (real parts).

– They constitute an *orthonormal basis* of $L^2(\mathbb{S}^2, \mathbb{C}; \sigma)$;

– They satisfy the *addition property*:

$$\frac{4\pi}{2n+1} \sum_{k=-n}^{+n} Y_{n,k}(x) \bar{Y}_{n,k}(y) = P_n(x \cdot y) \qquad x, y \in \mathbb{S}^2,\ n \in \mathbb{N}.$$

2 Background: Stochastic aspects

Let $Z = \big(Z(x), x \in \mathbb{S}^2\big)$ be a real-valued random field on $\mathbb{S}^2$. Z is said to be *second order stationary* (or *second order isotropic*) if, (i) there exists $m \in \mathbb{R}$ such that $E\{Z(x)\} = m$ for each $x \in \mathbb{S}^2$, and, (ii) there exists a function $C : [0, \pi] \mapsto \mathbb{R}$ such that $Cov\{Z(x), Z(y)\} = C\big(\alpha(x, y)\big)$ for each $x, y \in \mathbb{S}^2$. Here, m is the mean of the random field and C its covariance function. It was established by Schoenberg (1942) that C is a covariance function on $\mathbb{S}^2$ if and only if

$$C(\alpha) = \sum_{n=0}^{\infty} a_n P_n(\cos \alpha)$$

for some summable series (a_n) of non-negative terms. The measure $\sum_{n=0}^{\infty} a_n \delta_n$ is called the *spectral measure* (or angular power spectrum) of C.

Covariance functions on spheres may largely differ from those on Euclidean spaces. For instance, Gneiting [1] mentions that the function $C(\alpha) = \exp(-\alpha^2)$ is not positive definite on spheres. Starting from any two-dimensional Euclidean and isotropic covariance function C_e, and denoting its radial part by C_2 (i.e. $C_e(h) = C_2(|h|)$), Yadrenko [4] shows that the function $C(\alpha) = C_2\big(2\sin(\alpha/2)\big)$ is positive definite on $\mathbb{S}^2$. Although not general[1], the Yadrenko class is large enough to encompass all possible type of behaviours at the vicinity of the origin.

The random field is said to be Gaussian if any finite linear combination of its variables is Gaussianly distributed. As a Gaussian variable is characterized by its mean and its variance, the spatial distribution is characterized by its mean and its covariance function (or its spectral measure).

[1] For instance, the function $C(\alpha) = \lambda(1 - 2\alpha/\pi)$ is positive definite on $\mathbb{S}^2$, but not a a Yadrenko covariance function. It is the covariance function of the Chentsov-type random field $Z(x) = \sum_{p \in \mathcal{P}} \big(2\, 1_{\alpha(x,p) < \pi/2} - 1\big)$ where $\mathcal{P}$ is a homogeneous Poisson point process with intensity λ on $\mathbb{S}^2$.

3 Simulation of a Gaussian random field on $\mathbb{S}^2$

Our objective is to simulate a Gaussian random field with mean m and covariance function C on $\mathbb{S}^2$. Of course, there is no inconvenience in assuming the random field standardized (zero mean, unit variance), in which case the spectral measure $\sum_{n=0}^{\infty} a_n \delta_n$ is a probability measure, say $\mathcal{S}$.

Draw a random degree N from $\mathcal{S}$, a random order K uniform over $\{-N,...,+N\}$ and an independent phase Φ uniform over $[0,2\pi[$. Then consider the random field

$$Z(x) = 2\sqrt{2\pi}\, Re\big(Y_{N,K}(x)e^{i\Phi}\big) \qquad x \in \mathbb{S}^2.$$

The presence of the phase ensures that Z is centered. Regarding the covariance function, let us start with $Z(x) = \sqrt{2\pi}\left[Y_{N,K}(x)e^{i\Phi} + \bar{Y}_{N,K}(x)e^{-i\Phi}\right]$. Then we have

$$\begin{aligned} Cov\{Z(x),Z(y)\} &= 2\pi E\left\{\left[Y_{N,K}(x)e^{i\Phi} + \bar{Y}_{N,K}(x)e^{-i\Phi}\right]\left[\bar{Y}_{N,K}(y)e^{-i\Phi} + Y_{N,K}(y)e^{i\Phi}\right]\right\} \\ &= 2\pi E\left\{Y_{N,K}(x)\bar{Y}_{N,K}(y) + \bar{Y}_{N,K}(x)Y_{N,K}(y)\right\} \end{aligned}$$

because all terms with the random phase vanish when taking the expectation. Then, the *addition property* yields

$$Cov\{Z(x),Z(y)\} = E\{P_N(x\cdot y)\},$$

and it remains to calculate this expectation to obtain

$$Cov\{Z(x),Z(y)\} = \sum_{n=0}^{\infty} a_n P_n(x\cdot y) = \sum_{n=0}^{\infty} a_n P_n\big(\cos\alpha(x,y)\big) = C\big(\alpha(x,y)\big).$$

In particular, $Var\{Z(x)\} = \sum_{n=0}^{\infty} a_n = 1$. Accordingly, Z is standardized. Now, it should be pointed out that Z is not a Gaussian random field. However, if independent copies $Z_1,...,Z_p$ of the *basic random field* Z are generated, then the *Central Limit Theorem* states that the spatial distribution of $Z^{(p)} = (Z_1 + \cdots + Z_p)/\sqrt{p}$ tends to be multivariate Gaussian as p becomes very large. Among the possible criteria to select p, the 4^{th} order moment approach is particularly simple. Consider the linear combinations $Z(\Lambda) = \sum_{i=1}^{n} \lambda_i Z(x_i)$ and $Z^{(p)}(\Lambda) = \sum_{i=1}^{n} \lambda_i Z^{(p)}(x_i)$. Both variables $Z(\Lambda)$ and $Z^{(p)}(\Lambda)$ have the same variance σ^2 but different moments of order 4, denoted by μ_4 and $\mu_4^{(p)}$. If Z was multivariate Gaussian, then we would have $\mu_4 = \mu_4^{(p)} = 3\sigma^4$. A simple calculation shows

$$\mu_4^{(p)} - 3\sigma^4 = \frac{1}{p}\big(\mu_4 - 3\sigma^4\big).$$

4 Determination of the spectral measure

This simulation algorithm rests on the spectral measure of the covariance function. Literature says very little about it, up to the notable exception of Terdik [3].

Explicit calculations are sometimes possible. This is for instance the case of the Chentsov covariance function $C(\alpha) = 1 - 2\alpha/\pi$. The even coefficients of the spectral measure vanish. The odd ones are related by the induction formula

$$a_n = \frac{2n+1}{2n-3}\,\frac{(n-2)^2}{(n+1)^2}\,a_{n-2},$$

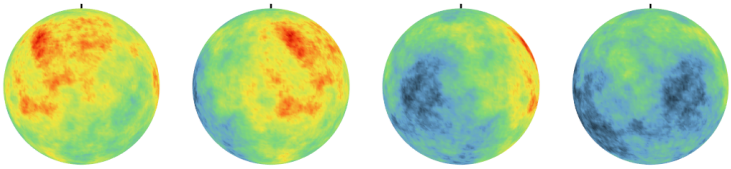


Figure 2: 4 views at 90° angles of a Gaussian random field with a Chentsov covariance function.

starting from $a_1 = 3/4$. The simulation of Fig. 2 has been obtained using 15000 basic random fields.

Another example is the exponential covariance function $C(\alpha) = e^{-\nu\alpha}$ where $\nu > 0$. The even and odd coefficients of the spectral measure satisfy the same induction formula

$$a_n = \frac{2n+1}{2n-3}\,\frac{\nu^2+(n-2)^2}{\nu^2+(n+1)^2}\,a_{n-2},$$

starting from $a_0 = (1+e^{-\nu\pi})/(1+\nu^2)$ and $a_1 = 3(1-e^{-\nu\pi})/(4+\nu^2)$. Fig. 3 shows a simulation with $\nu = 0.1$. 20000 basic random fields were used.

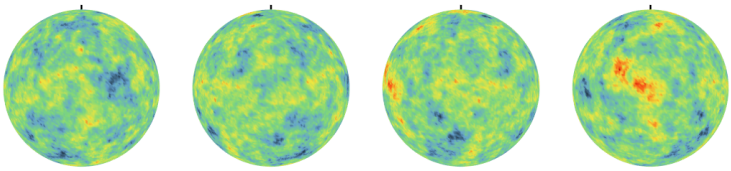


Figure 3: 4 views at 90° angles of a Gaussian random field with a exponential covariance function.

In the case of Yadrenko covariance functions, the spectral measure of C can be related to that of C_e (as provided by Bochner theorem). Denoting by $dF_2(r)$ its radial part, the following 3 formulae can be used to calculate or compute the a_n's (J_λ is the Bessel function of order λ):

$$a_n = \frac{2n+1}{2}\int_{-1}^{+1} C_s(\arccos t)\,P_n(t)\,dt,$$
$$a_n = \frac{2n+1}{2}\int_{-1}^{+1} C_2\big(\sqrt{2(1-t)}\big)\,P_n(t)\,dt,$$
$$a_n = 2\pi(2n+1)\int_0^{\infty} J_{2n+1}(2r)\,dF_2(r).$$

Acknowledgements: This work was supported by the project ANR-15-ASTR-0024. The author is grateful to E. Chautru for her careful reading of this paper.

References

[1] Gneiting, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli* **19-4**, 1327–1349.

[2] Marinucci, D. and Peccati, G. (2011). *Random fields on the sphere*. Cambridge University Press.

[3] Terdik, G. (2015). Angular spectra for non-Gaussian isotropic fields. *Braz. J. Prob. Stat.* **29-4**, 833–865.

[4] Yadrenko, M.I. (1983). *Spectral Theory of random fields*. Springer-Verlag. New York.

Axial Symmetry Models for Space-Time Data

Emilio Porcu[1,*], Alfredo Alegria[1] Stefano Castruccio[2] and Paola Crippa[3]

1 *School of Mathematics, Statistics and Physics, Newcastle University, UK; georgepolya01@gmail.com, alfredo.alegria@gmail.com*
[2] *Department of Applied and Computational Mathematics and Statistics, University of Notre Dame; scastruc@nd.edu*
[3] *Department of Civil & Environmental Engineering & Earth Sciences, University of Notre Dame; pcrippa@nd.edu* *Corresponding author*

Abstract. *This paper revisits axial symmetry for spatial data, putting special emphasis on results obtained recently by the same group of authors. Then, axial symmetry for space-time data is discussed and new models for this characterization are proposed. We illustrate our findings on applications to climate models.*

Keywords. *Axial Symmetry; Covariance Function; Space-Time*

---

1 Introduction

The increasing availability of data on a global scale, both from climate model simulations and from satellite observations, calls for providing statistical methodologies that are suitable for spherical domains. The construction of valid global models is, however, very different from the Euclidean geometries, being very popular in classical geostatistics. Even with Gaussian processes, covariance functions over spherical domains require a very different mathematical theory (Gneiting, 2013; Berg and Porcu, 2017; Jeong et al., 2018; Porcu et al., 2018a). Furthermore, while for Euclidean data the isotropic assumption might be suitable (at least as a first-order approximation), and could be the building block for the development of more sophisticated models, this is not the case for global data. As explained by Porcu et al. (2018b), *isotropy might be suitable for microscale meteorology, but it is not for mesoscale and synoptic scale meteorology, where the Earth's axis inclination, global circulation and teleconnections generate dependencies in dominant directions for many physical variables*. While statistical models must be tailored for a given variable and account for the physics of the problem, a sensible first-order approximation would be that of heterogeneous spatial dependence across latitude, while stationarity could be assumed across longitude. This class of models, called *axially symmetric* (Jones, 1963), has been proposed as a standard for global data by Stein (2007) while analyzing total Ozone mapping spectrometer data. Different approaches have been developed to allow closed form expression of covariance functions via partial derivatives (Jun and Stein, 2007, 2008), spectral characterizations (Hitczenko and Stein, 2012), as well as fast spectral inference schemes for massive datasets (Castruccio and Stein, 2013; Castruccio and Genton, 2018). Recently, Porcu et al. (2018b) have proposed an axially symmetric version of the Matérn model, where both scale and smoothness parameters are adapted to become functions of latitudes.

This paper revisits axial symmetry for spatial processes on spheres and discusses the construction of

space-time processes that are spatially axially symmetric.

2 Axial Symmetry

Let $\mathbb{S}^2 = \{\mathbf{s} \in \mathbb{R}^3, \|\mathbf{s}\| = 1\}$, be the unit sphere, where $\|\cdot\|$ denotes the Euclidean distance. Any point $\mathbf{s} \in \mathbb{S}^2$ is represented through its spherical coordinates $\mathbf{s} = (L, \ell)$, with $L \in [-\pi/2, \pi/2]$ and $\ell \in [-\pi, \pi)$ being respectively the latitude and longitude (equivalently, polar and the azimuthal angles).

The geodesic is defined as the mapping $d_{\mathrm{GC}} : \mathbb{S}^2 \times \mathbb{S}^2 \to [0, \pi]$ so that

$$d_{\mathrm{GC}}(\mathbf{s}_1, \mathbf{s}_2) = \arccos\big(\langle \mathbf{s}_1, \mathbf{s}_2 \rangle\big) = \arccos\big(\sin L_1 \sin L_2 + \cos L_1 \cos L_2 \cos|\Delta\ell|\big),$$

with $\mathbf{s}_i = (L_i, \ell_i)$, $i = 1, 2$, and $\langle \cdot, \cdot \rangle$ denoting the dot product on the sphere, and where $\Delta\ell = \ell_1 - \ell_2$. Henceforth, we shall equivalently use $d_{\mathrm{GC}}(\mathbf{s}_1, \mathbf{s}_2)$ or the shortcut d_{GC} to denote the geodesic distance, whenever no confusion arises. which defines a segment below the arc joining two points on the spherical shell.

Consider a zero mean Gaussian random field over the sphere $\{Z(\mathbf{s}),\ \mathbf{s} \in \mathbb{S}^2\}$ with finite second order moment. The finite dimensional distributions are therefore completely specified by the covariance function $C : \mathbb{S}^2 \times \mathbb{S}^2 \to \mathbb{R}$, defined by

$$C(\mathbf{s}_1, \mathbf{s}_2) = \mathbb{C}\mathrm{ov}\big(Z(\mathbf{s}_1), Z(\mathbf{s}_2)\big), \qquad \mathbf{s}_1, \mathbf{s}_2 \in \mathbb{S}^2.$$

Covariance functions are positive definite: for any κ dimensional collection of points $\{\mathbf{s}_i\}_{i=1}^{\kappa} \subset \mathbb{S}^2$ and constants $c_1, \ldots, c_\kappa \in \mathbb{R}$, we have $\sum_{i=1}^{\kappa} \sum_{j=1}^{\kappa} c_i C(\mathbf{s}_i, \mathbf{s}_j) c_j \geq 0$, see Bingham (1973). Porcu et al. (2018a) call C geodesically isotropic if

$$C(\mathbf{s}_1, \mathbf{s}_2) = \psi(d_{\mathrm{GC}}(\mathbf{s}_1, \mathbf{s}_2)), \tag{1}$$

for some $\psi : [0, \pi] \to \mathbb{R}$. The function ψ is called the geodesically isotropic part of C (Daley and Porcu, 2013). For a characterization of geodesic isotropy, the reader is referred to Schoenberg (1942), the essay in Gneiting (2013) and the more recent work of Berg and Porcu (2017).

For quantities observed on a global scale, isotropy is not tenable. While processes at small scale (micro-scale, turbulence scale) might be approximately regarded as isotropic, large-scale meteorological patterns have preferred directions driven by general circulation. Indeed, Stein (2007) showed that total column ozone data show significant changes over latitudes. Castruccio and Stein (2013) argued that both the inter- and intra-annual variability for surface temperature are dependent on latitude. This leads to the definition of an *axially symmetric* covariance C when

$$C(\mathbf{s}_1, \mathbf{s}_2) = \mathcal{C}(L_1, L_2, \Delta\ell). \tag{2}$$

for some function $\mathcal{C} : [-\pi/2, \pi/2]^2 \times [-2\pi, 2\pi] \to \mathbb{R}$.

The Matérn function $\mathcal{M}_\nu(\cdot; \alpha) : [0, \infty) \to [0, \infty)$ is defined as

$$\mathcal{M}_\nu(d; \alpha) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{d}{\alpha}\right)^{\nu} \mathcal{K}_\nu\left(\frac{d}{\alpha}\right), \qquad d \geq 0, \tag{3}$$

where $\alpha > 0$ is the range, and $\nu > 0$ is the parameter that controls the mean square differentiability of the process (Stein, 1999). Observe that we do not call it *covariance* function because to become a covariance it must be composed with a metric. In classic application people use the Euclidean distance on the plane, or the chordal distance on the sphere. In fact, $\mathcal{M}_\nu(d_{\mathrm{GC}};\alpha)$ is no longer a covariance function, unless $\nu \in (0,1/2]$ (Gneiting, 2013).
Porcu et al. (2018b) propose a Matérn version that is axially symmetric. Specifically, they have

$$\mathcal{C}(L_1,L_2,\ell) = \sigma(L_1,L_2)\mathcal{M}_{\widetilde{\nu}(L_1,L_2)}\left(|\Delta\ell|;\sqrt{\widetilde{\alpha}(L_1,L_2)}\right), \qquad (L_1,L_2,\Delta\ell) \in [-\pi/2,\pi/2]^2 \times [-2\pi,2\pi], \quad (4)$$

where the continuous functions $\sigma, \widetilde{\alpha}, \widetilde{\nu}$ must respect some constraint that we do not specify here to avoid mathematical obfuscation.

2.1 Space-Time and Axial Symmetry

We now consider a zero mean Gaussian random field over the sphere cross time, that is $\{Z(\mathbf{s},t),\ \mathbf{s} \in \mathbb{S}^2,\ t \in \mathbb{R}\}$ with finite second order moment. The finite dimensional distributions are therefore completely specified by the covariance function $C : \mathbb{S}^2 \times \mathbb{R} \times \mathbb{S}^2 \times \mathbb{R} \to \mathbb{R}$, defined by

$$C\big((\mathbf{s}_1,t_1),(\mathbf{s}_2,t_2)\big) = \mathbb{C}\mathrm{ov}\big(Z(\mathbf{s}_1,t_1),Z(\mathbf{s}_2,t_2)\big), \qquad \mathbf{s}_1,\mathbf{s}_2 \in \mathbb{S}^2, \quad t_1,t_2 \in \mathbb{R}.$$

The function C is called separable if it factors into the product of a merely spatial with a merely temporal covariance. We call the function C axially symmetric and temporally stationary if

$$C\left((\mathbf{s}_1,t_1),(\mathbf{s}_2,t_2)\right) = \mathcal{C}(L_1,L_2,\Delta\ell,\Delta t). \qquad (5)$$

for some function $\mathcal{C} : [-\pi/2,\pi/2]^2 \times [-2\pi,2\pi] \times \mathbb{R} \to \mathbb{R}$. Here, $\Delta t = t_1 - t_2$. Characterization of the functions $\mathcal{C}$ that satisfy this hypothesis has been elusive so far. To the knowledge of the authors, there is a clear lack of nonseparable parametric models that allow to model this kind of situation. This will be the central core of our proposal. We give here an informal suggestion to avoid mathematical obfuscation. We shall consider the class

$$\mathcal{C}(L_1,L_2,\Delta\ell,\Delta t) = \frac{1}{f(L_1,L_2,\Delta t)^{1/2}} g\left(\frac{|\Delta\ell|}{f(L_1,L_2,\Delta t)}\right),$$

for two functions f and g that must be determined to ensure positive definiteness. We shall show that the assumption on the function f is very simple (for instance, f might be completely monotonic on the positive real line). Instead, some technical assumptions will be needed on the function g and these will be explained through the presentation at METMA conference.

References

Berg, C. and Porcu, E. (2017). From Schoenberg Coefficients to Schoenberg Functions. *Constructive Approximation*, 45:217–241.

Bingham, N. H. (1973). Positive Definite Functions on Spheres. *Mathematical Proceedings of the Cambridge Philosophical Society*, 73:145–156.

Castruccio, S. and Genton, M. G. (2018). Principles for Inference on Big Spatio-Temporal Data from Climate Models. *Statistics and Probability Letters*. in press.

Castruccio, S. and Stein, M. L. (2013). Global Space-Time Models for Climate Ensembles. *Annals of Applied Statistics*, 7(3):1593–1611.

Daley, D. J. and Porcu, E. (2013). Dimension Walks and Schoenberg Spectral Measures. *Proceerings of the American Mathematical Society*, 141:1813–1824.

Gneiting, T. (2013). Strictly and Non-Strictly Positive Definite Functions on Spheres. *Bernoulli*, 19(4):1327–1349.

Hitczenko, M. and Stein, M. L. (2012). Some Theory for Anisotropic Processes on the Sphere. *Statist. Methodology*, 9:211–227.

Jeong, J., Jun, M., and Genton, M. (2018). Covariance Models for Global Spatial Statistics. *Statistical Science*, To Appear.

Jones, R. H. (1963). Stochastic Processes on a Sphere. *Annals of Mathematical Statistics*, 34:213–218.

Jun, M. and Stein, M. L. (2007). An Approach to Producing Space-Time Covariance Functions on Spheres. *Technometrics*, 49:468–479.

Jun, M. and Stein, M. L. (2008). Nonstationary Covariance Models for Global Data. *Annals of Applied Statistics*, 2(4):1271–1289.

Porcu, E., Alegría, A., and Furrer, R. (2018a). Modeling Temporally Evolving and Spatially Globally Dependent Data. *International Statistical Review*. in press.

Porcu, E., Castruccio, S., Alegría, A., and Crippa, P. (2018b). Axially symmetric models for global data: a journey between geostatistics and stochastic generators. *Submitted*.

Schoenberg, I. J. (1942). Positive Definite Functions on Spheres. *Duke Mathematical Journal*, 9:96–108.

Stein, M. L. (1999). *Interpolation of Spatial Data. Some Theory of Kriging*. Springer-Verlag, New York.

Stein, M. L. (2007). Spatial Variation of Total Column Ozone on a Global Scale. *Annals of Applied Statistics*, 1(1):191–210.

Contours and dimple for the Gneiting class of space-time correlation functions

Francisco Cuevas[1,*], Emilio Porcu[2] and Moreno Bevilacqua[3]

[1] *Department of Mathematical Sciences, Aalborg University, Skjernvej 4A, 9220, Denmark; francisco@math.aau.dk*
[2] *School of Mathematics and Statistics, University of Newcastle, Newcastle upon Tyne NE1 7RU, UK; georgepolya01@gmail.com*
[3] *Department of Statistics, University of Valparaiso, Avenida Gran Bretaña 1091, Chile; moreno.bevilacqua@uv.cl*
[*] *Corresponding author*

Abstract. *We offer a dual view of the dimple problem related to space-time correlation functions in terms of their contours. We find that the dimple property (see [4]) in the Gneiting class of correlations is in one-to-one correspondence with non-monotonicity of the parametric curve describing the associated contour lines. Further, we show that, given such a non monotonic parametric curve associated to a given level set, all the other parametric curves at smaller levels inherit the property of non monotonicity. We finally propose a modified Gneiting class of correlations having monotonically decreasing parametric curves and no dimple along the temporal axis.*

Keywords. *Covariance function; Dimple; Gneiting class.*

1 Introduction

Spatio-temporal geostatistics deals mainly with the second order properties of stochastic processes defined over a spatial domain and evolving over time. In particular, covariance and correlation functions allow one to describe the interactions between spatial and temporal components, and they are crucial to both estimation and prediction [5].

Let $Z(x,u)$ denote a space-time random field with continuous spatial index $x \in \mathbb{R}^d$ and temporal index $u \in \mathbb{R}$. This work is concerned with stationary correlation functions that are spatially isotropic and temporally symmetric, i.e.,

$$\operatorname{corr}\{Z(x,u), Z(x+h,u+v)\} = C(\|h\|,|v|), \qquad (h,v) \in \mathbb{R}^d \times \mathbb{R}, \tag{1}$$

with C being continuous and such that $C(0,0) = 1$, and where $\|\cdot\|$ denotes the Euclidean norm. Through this work we abuse of terminology and we say that C is the space-time correlation function of Z. Also, we write (r,t) for the pair $(\|h\|,|v|)$ above. [2] provide the so called Gneiting class of correlation functions, given by

$$C(r,t) = \frac{1}{\psi(t^2)^\delta}\varphi\left\{\frac{r^2}{\psi(t^2)}\right\}, \qquad (r,t) \in [0,\infty)\times[0,\infty), \tag{2}$$

where φ is a completely monotone function and ψ is a Bernstein function. In what follows we fix δ and without loss of generality we assume that $\psi(0) = \varphi(0) = 1$. In addition, we assume that $\varphi(t) \to 0$ when $t \to \infty$ and that $\psi(t) \to \infty$ when $t \to \infty$.

Gneiting's class of correlation functions are very flexible. However, depending on the choice of φ and ψ in Equation (2), the Gneiting class can be counterintuitive. Such problem was detailed in [4] and was called *the dimple property*. Dimple in a space-time correlation can be understood as follows: $Z(x_{\text{here}}, u_{\text{now}})$ is more correlated with $Z(x_{\text{there}}, u_{\text{tomorrow}})$ than with $Z(x_{\text{there}}, u_{\text{now}})$. Such property should be taken into account when modeling space-time data ([4]).

In this work we study the contours $\mathcal{C}_y$ of Equation (2) to describe the dimple property. In this case $\mathcal{C}_y$ is described by the parametric curve $r = f(t;y)$ which takes the form

$$f(t;y) = \left[\varphi^{-1}\left\{y\psi(t^2)^\delta\right\}\psi(t^2)\right]^{1/2}, \qquad y \in (0,1), t > 0, \tag{3}$$

where t is such that $y\psi(t^2)^\delta \leq 1$, so that f is well-defined. Here, φ^{-1} denotes the proper inverse of φ. In particular, $\varphi^{-1}(1) = 0$ and $\varphi^{-1}(0) = \infty$.

In this work we show that the dimple property of the Gneiting class in Equation (2) is intimately related to a non-monotonic property of the parametric curves depicted through Equation (3). Further, we show that for any fixed y^* such that $f(\cdot;y^*)$ is non-monotonically decreasing, all the curves of the type $f(\cdot;y^*-\varepsilon)$, for all $\varepsilon > 0$, will be non-monotonically decreasing as well. Motivated by the constructive criticism on the counterintuitive behavior of correlation functions that satisfy the dimple property, we propose a correlation function of the Gneiting type but with no dimple along the temporal axis. The proofs of the theoretical results are deferred to [1].

2 Dimple, contours, and parametric curves

The dimple property appears when, for a fixed spatial lag r_0, the temporal margin $C(r_0,\cdot)$ is non-monotonically decreasing. The formal statement was introduced by [4] and is detailed below.

Definition A stationary, spatially isotropic and temporally symmetric correlation function $C(r,t)$ has a dimple along the time lag t if there exists a $z^* > 0$ such that the following properties hold:

(a) for fixed $r^2 \leq z^*, C(r,t)$ is decreasing in $t \geq 0$;

(b) for fixed $r^2 > z^*, C(r,t)$ is increasing for $t \in (0,t^*)$ for some $t^* = t^*(r^2) > 0$, and decreasing for $t \in (t^*,\infty)$.

We now follow [4] when introducing the function $Q : \mathbb{R}_+ \to [0,\infty]$, defined through

$$Q(z) = -\frac{z\varphi'(z)}{\varphi(z)}, \quad z > 0. \tag{4}$$

The function Q is the crux of the following criterion (provided by [4]).

Theorem Consider Gneiting's model as defined through Equation (2). If the function $Q(z)$ defined by (4) is increasing in $z > 0$ and $\delta < \lim_{z\to\infty} Q(z) \leq \infty$, then Gneiting's correlation function has a dimple along the time lag t.

To simplify notation, we write

$$\left.\frac{\partial C(r,t)}{\partial r}\right|_{(r,t)=(r_0,t_0)} = C_r(r_0,t_0).$$

Analogously, we use C_t, f_t and f_r for the partial derivatives associated to C and f, if they exists, respectively. Our first finding is the following result.

Theorem Let C be the Gneiting correlation defined through Equation (2). Let f be given by Equation (3) and suppose that C_r is strictly positive on $[0,\infty)\times[0,\infty)$. Then, C has a dimple if and only if there exists a $y^* \in (0,1)$ such that $f(\cdot;y^*)$ has a maximum $r^* = f(t^*;y^*)$, with t^* from Definition 2.

For the Gneiting class of correlation functions, Theorem 2 gives an explicit relationship between the dimple effect and the non-monotonicity of the parametric curves f in Equation (3). Rephrased, we have that the Gneiting correlation C has a dimple if and only if there exists a $y^* \in (0,1)$ such that $f(\cdot;y^*)$ is non-monotonically decreasing in t.

Theorem Let $y^* \in (0,1)$ and $t^* > 0$ such that $f_t(t^*;y^*) = 0$. Then, for every $\varepsilon > 0$ such that $y^* - \varepsilon > 0$ there exists $\lambda > 0$ such that $f_t(t^* + \lambda; y^* - \varepsilon) = 0$.

To further clarify, an example is given. Consider Equation (4) in [2], which is obtained by choosing $\varphi(t) = \exp(-t^\alpha), \psi(t) = (1+t^\gamma)^\beta$ in Equation (2). We consider the special case $\alpha = \gamma = 1/2$ to obtain

$$C(r,t) = \exp\left\{-\frac{r}{(1+t)^{\beta/2}}\right\}(1+t)^{-\beta\delta}, \quad 0 < \beta \leq 1, \quad \delta > d/2 \quad r,t > 0. \tag{5}$$

The associated contour lines assume the form

$$f(t;y) = \{-\log y - \beta\delta\log(1+t)\}(1+t)^{\beta/2}, \quad t > 0, y \in (0,1). \tag{6}$$

Direct inspection of the zeros of the first derivative shows that $f_t(t;y)$ is identically equal to zero at a given $y \in (0,1)$ for $t = -1 + \exp(-2/\beta)y^{-1/(\beta\delta)}$. Such a zero is located on the positive real line if and only if $-\log y \geq 2\delta \geq d$. This gives a direct interpretation of the dimple in terms of dimension d and contour levels y.

3 A Gneiting class with monotonic parametric curves and no dimples

In [4] the dimple is said to be counterintuitive, so it would be desirable to modify the Gneiting class in order to have a new correlation with no dimple along the temporal axis. The following result describes how to obtain such a modified Gneiting class.

Theorem Let $\delta \geq (d+3)/2$ and F be a nonnegative finite Borel measure on $[0,\infty)$ such that $\int_0^\infty s^\delta F(s) = 1$. Let

$$\varphi(x) = \int_0^\infty \exp(-xs)s^\delta F(s), \quad x \geq 0.$$

Also, let ψ be continuous on $[0,\infty)$, and an increasing and concave function on the positive real line with $\psi(0) = 1$. Then

$$C(r,t) = \frac{\varphi\{\psi(t)r\}}{\psi(t)^{\delta}}, \quad (r,t) \in [0,\infty) \times [0,\infty), \tag{7}$$

is a correlation function on $\mathbb{R}^d \times \mathbb{R}$ that is radially symmetric in the spatial argument and symmetric in the temporal one.

The parametric curve $\widetilde{f}$ associated to C in Equation (7) has expression

$$\widetilde{f}(t;y) = \frac{\varphi^{-1}\left\{y\psi(t)^{\delta}\right\}}{\psi(t)}, \quad y \in (0,1), t \geq 0, \tag{8}$$

where $y\psi(t)^{\delta} \leq 1$.

In the Appendix we show that $\widetilde{f}(\cdot;y)$ is strictly decreasing for any fixed $y \in (0,1)$. One may note how the non-dimple property in the modified Gneiting class in Equation (7) is achieved at the expense of a more severe restriction on δ.

Consider the measure $F(\mathrm{d}s) = s^{-1}\exp(-s)/\Gamma(\delta)\mathrm{d}s$, for $\delta \geq (d+3)/2$ and Γ denoting the Gamma function. Further, with

$$\varphi(x) = \int_0^{\infty} \exp(-xs)s^{\delta}F(\mathrm{d}s) = (1+x)^{-\delta}, \qquad x > 0,$$

and $\psi(t) = (1+t)^{\alpha}$, for $\alpha \in (0,1]$, we obtain

$$C(r,t) = \{1 + r(1+t)^{\alpha}\}^{-\delta}(1+t)^{-\alpha\delta}, \quad r,t \geq 0.$$

The associated parametric curve is strictly decreasing for any $y \in (0,1)$.

Acknowledgments. The first author would like to thank the MSH (Maison des Sciences de l'Homme) and the Department of Mathematics of the University of Montpellier, as well as SFdS and CNRS for supporting the METMA IX workshop.

References

[1] Cuevas, F., Porcu, E., Bevilacqua, M. (2017). Contours and dimple for the Gneiting class of space-time correlation functions, *Biometrika*, **4**, 995–1001

[2] Gneiting , T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, **97**, 590–600.

[3] Gneiting, T., Genton, M., Guttorp, P. (2007). Geostatistical space-time models, stationarity, separability and full symmetry. *Monographs in Statistics and Applied Probability, Chapman & Hall/CRC Press*, 151–75.

[4] Kent, J. T., M. Mohammadzadeh, Mosammam, A. M. (2011). The dimple in Gneiting's spatial-temporal covariance model, *Biometrika*, **98**, 489–494.

[5] Stein, M. L. (2005). Space-time covariance functions. *Journal of the American Statistical Association*, **100**, 310–321.

Space-time extreme processes simulation for flash floods in Mediterranean France

F. Palacios-Rodríguez[1,*], G. Toulemonde[1] , J. Carreau[2] and T. Opitz[3]

[1] *Institut Montpelliérain Alexander Grothendieck. Université de Montpellier. Place Eugène Bataillon 34095 Montpellier, France; gwladys.toulemonde@umontpellier.fr, fatima.palacios-rodriguez@umontpellier.fr*
[2] *HydroSciences Montpellier (UMR 5569). Université Montpellier. 163 Rue Auguste Broussonet 34090 Montpellier, France; julie.carreau@ird.fr*
[3] *BioSP, INRA, 228 Route de l'Aérodrome, 84914 Avignon Cedex 9; thomas.opitz@inra.fr*
[*] *Corresponding author*

Abstract. *Flash floods in France are highly destructive natural phenomena, not only by creating material damage but also by threatening the safety of human life. To anticipate the impact of such disasters, it is crucial to propose stochastic simulation methods of realistic scenarios for spatio-temporal extreme fields. Pareto processes are justified because they model phenomena that exceed a certain extreme threshold. Therefore, they are promising models for the aforementioned challenge. Nonparametric and parametric approaches in this framework have been provided over last years, but the proposed models do not establish a direct link to Pareto processes. A semiparametric method for simulation of extreme space-time generalized Pareto processes is introduced. A key benefit of the proposed method is that it allows to generate an unlimited number of realizations of such extreme fields. Our simulation method is applied to a rainfall data-set to model flash floods in a region in Mediterranean France.*

Keywords. *Environmental risk; Extreme Value modelling; Pareto processes; Simulation; Space-time processes.*

---

1 Introduction

Flash floods can be characterized as floods with a sudden and fast rise of stream flow, and a large peak flow in terms of specific discharge rate. These floods are linked to intense local rainfall produced in a short time period or to longer rains with moderate intensities that affect the entire catchment area ([6]). Over the last 25 years, flash floods in France have constituted one of the most destructive natural phenomena not only creating material damage but also threatening the safety of human life ([9]). Therefore, the understanding of temporal and spatial variability of the rainfall patterns that generate these floods has received considerable attention by the authorities. To help with this comprehension through the analysis of impact models fed with a large number of potential precipitation scenarios, the construction of stochastic simulation methods of scenarios incorporating realistic spatio-temporal extreme fields is crucial. To deal with this challenge, we mobilize statistical techniques based on Extreme Value Theory ([3]).

Several nonparametric and parametric approches have been developed for stochastic simulation of extreme fields ([10] and [4]). Recently, [2] proposed a semiparametric method to simulate extreme spatio-temporal fields of wave heights in the Gulf of Lions (France) based on the procedure described in [1]. Although [2]'s approach provides an appropriate simulation technique for wave heights in the Gulf of Lions, several of its aspects call for methodological improvements. First, this procedure implies

that the numbers of possible simulations has to be limited is defined in [2]. Second, Pareto processes are the natural models for this problem since they model phenomena that exceed a certain threshold ([7]). However, the method in [2] does not establish this link to Pareto processes.

The purpose of this work is to provide a generalization of [2]'s approach to overcome the two outlined drawbacks (see Section 3). By using such spatio-temporal extreme modelling, we aim to extract information from rainfall data in a mediterranean region in France to obtain extremes simulations of the event, potentially more extreme than those from the observation period. The data-set is described in Section 2.

2 Rainfall reanalysis data-set for Mediterranean France

From an applied perspective, we aim to create realistic extreme simulations for flash floods in a region in Mediterranean France. There are different stations localized in our study region where the rainfall data are measured by irregularly spaced rain gauges. However, in order to develop our study taking into account the space variable in a more continuous way, the rainfall measurements have to be available over a sufficiently dense grid of sites. Therefore, the considered rainfall reanalysis data-set was constructed by a combination of two strategies: on the one hand, algorithms that process radar signals (remotely sensed rainfall measurements), and on the other hand, water slides of rain gauges. The reanalysed data-set contains hourly rainfall measurements recorded at 10914 sites that cover a grid of resolution $1km^2$ in Mediterranean France, from 1997 to 2007. The unit of measurement is *mm*. The maximum precipitation value observed over time at each site is presented in Figure [1] (left panel). The data-set considered was provided by *Météo-France* (http://www.meteofrance.com).

A preliminary analysis is realised to improve our understanding of the rainfall data-set. Since our method is closely related to max-stable modeling (see Equation (1)), we quantify the asymptotic bivariate extremal dependence by using the extremal coefficient θ, $1 \leq \theta \leq 2$. The limiting case $\theta = 1$ represents full dependence, whereas $\theta = 2$ represents independence. In our case, we study extremal coefficients for the spatial and temporal dimension. The spatial extremal coefficient $\theta^{spa}(h)$ measures the dependence between pairs of rain data separated by a spatial distance h, at a given time. The time extremal coefficient $\theta^{tim}(k)$ measures the dependence between pairs of rain data separated by a time lag k, at a given site (see Section 2.2 in [2] for more details). Figure [1] (center panel) shows that $\hat{\theta}^{spa}(h)$ is always clearly lower than 2. In addition, from Figure [1] (right panel), we can conclude that the duration of a storm in our region is not over 2 days. From these observations, it follows that using max-stable-based modeling techniques, which are useful for asymptotically dependent data, is appropriate.

The previous analysis for the data-set as well as the proposed algorithm in Section 3 are implemented in parallellized R code.

3 Methodology

In this section, we briefly outline our novel semiparametric spatio-temporal extreme model, which behaves asymptotically as a generalized Pareto Process ([5]).

Let $C(\mathcal{S} \times \mathcal{T})$ be the space of continuous real functions on $\mathcal{S} \times \mathcal{T}$, equipped with the supremum norm, where $\mathcal{S}$ is a compact subset of $\mathbb{R}^2$ and $\mathcal{T}$ is a compact subset of $\mathbb{R}^+$. Consider a continuous stochastic process $\{X(s,t)\}_{s\in\mathcal{S},t\in\mathcal{T}} \in C(\mathcal{S} \times \mathcal{T})$. Suppose that the probability distribution of the process is in the domain of attraction of some max-stable process, that is, there exist functions $a_n > 0$ and b_n such that the

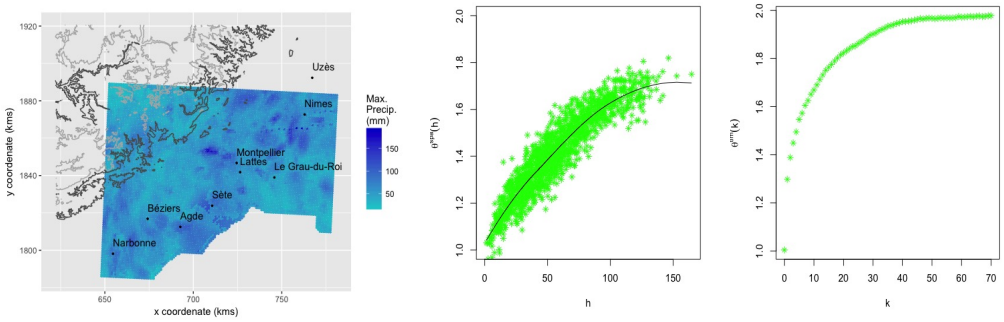


Figure 1: Left panel: Maximum precipitations by sites in a mediterranean region in France. Center panel: $\hat{\theta}^{spa}(h)$ associated to pairs of sites separated by 1500 different distances h. The observation period is October 1997-April 1998. Fitted local polynomial regression model (black line). Right panel: $\hat{\theta}^{tim}(k)$ for rain pairs separated by a time lag k.

sequence of i.i.d. processes

$$\left\{ \max_{1\leq i\leq n} \frac{X_i(s,t)-b_n(s,t)}{a_n(s,t)} \right\}_{s\in\mathcal{S},t\in\mathcal{T}} \tag{1}$$

converges to a continuous process, say $Z(s,t)$, in distribution in $C(\mathcal{S}\times\mathcal{T})$. $Z(s,t)$ is called max-stable process. Let $\ell : C^+(\mathcal{S}\times\mathcal{T}) \to [0,+\infty)$ be a continuous nonnegative and homogeneous function, called *risk functional*. Let G be a distribution that belongs to the standard Fréchet maximum domain of attraction. Let G^{-1} be the inverse of G. Let $F_{X(s,t)}$ denote the distribution function of $X(s,t)$. We define a standard transformation $\tilde{T}$ as $\tilde{T}(X(s,t)) := G^{-1}\left(F_{X(s,t)}(X(s,t))\right)$, $\forall s\in\mathcal{S}$, $\forall t\in\mathcal{T}$. Following [8], it is convenient to fix a high threshold function $u(s,t)$ and assume that the marginal distribution satisfies

$$P(X(s,t)>x) = \left[1+\gamma(s,t)\frac{x-\mu(s,t)}{\sigma(s,t)}\right]_+^{-1/\gamma(s,t)} \tag{2}$$

for $x > u(s,t)$, with real parameters $\mu(s,t) < u(s,t)$, $\sigma(s,t) > 0$ and $\gamma(s,t)$, such that the right-hand side of (2) is less that unity.

The following simulation method for a generalized Pareto process $\{W_l^*(s,t)\}_{s\in\mathcal{S},t\in\mathcal{T}}$ is proposed:

1. Estimate $\gamma(s,t)$, $\sigma(s,t)$ and $\mu(s,t)$ in (2). Denote the estimators by $\hat{\gamma}(s,t)$, $\hat{\sigma}(s,t)$ and $\hat{\mu}(s,t)$, respectively. Denote by $\hat{\tilde{T}}X(s,t)$ and $\hat{\tilde{T}}^{\leftarrow}f(s,t)$ the expressions $\tilde{T}X(s,t)$ and $\tilde{T}^{\leftarrow}f(s,t)$ with $\hat{\gamma}(s,t)$, $\hat{\sigma}(s,t)$ and $\hat{\mu}(s,t)$ inside.[1]
2. Select for the normalized processes: $\frac{\hat{\tilde{T}}X_i(s,t)}{\ell(\hat{\tilde{T}}X_i(s,t))}$, $i=1,\ldots,n$, those that satisfy $\ell(\hat{\tilde{T}}X_i(s,t))>1$, for some $s\in\mathcal{S}$ and $t\in\mathcal{T}$.
3. Let R be a Pareto random variable with shape parameter 1 and scale parameter $\alpha>0$, that is, $P(R>x)=\alpha/x$, for $x\in[\alpha,+\infty)$. Generate $R_1,\ldots,R_n$ such that $R_i \overset{d}{=} R$, for $i=1,\ldots,n$.
4. Finally, generate $W^*_{\ell,i}(s,t) := \hat{\tilde{T}}^{\leftarrow}(W_{\ell,i}(s,t))$, where $W_{\ell,i}(s,t) := \left(R_i\frac{\hat{\tilde{T}}X_i(s,t)}{\ell(\hat{\tilde{T}}X_i(s,t))}\right)$, $i=1,\ldots,n$.

If we rewrite $W^*_\ell(s,t) := \tilde{T}^{\leftarrow}(W_\ell(s,t))$, where $W_\ell(s,t) := \left(R\frac{\tilde{T}X(s,t)}{\ell(\tilde{T}X(s,t))}\right)$, $\quad\forall s\in\mathcal{S},\forall t\in\mathcal{T}$, we can formally

[1] To develop the estimation procedure, we could consider the threshold and the parameters to be constant over time, depending only on space ([2]). In addition, we use the maximum likelihood estimation procedure applied to (2).

show that W_ℓ^* behaves asymptotically as a generalized Pareto process.

4 Conclusions

A semiparametric method for simulation of extreme space-time generalized Pareto processes is provided in this work. In contrast to the approach in [2], our method is not restricted to a limited number of extreme field simulations. Regarding the data transformation, our method (see item 2 in Section 3) can deal with more general transformation functions $\tilde{T}$ than the one proposed in [2]. In addition, we generalize the selection procedure of the extreme event by allowing for more general risk functionals ℓ (see item 2 in Section 3) than the maximum. While [2] consider a deterministic coefficient $\xi > 1$ for the uplifting procedure to generate more extreme, events of yet unobserved magnitude, we use realizations from a Pareto distribution with scale parameter α. This allows us to find nice interpretations in terms of α. The proposed method is applied to a rainfall data-set for a region in Mediterranean France in order to model flash floods in the region. Using our method, the rainfall process can be "uplifted", and we can consider a more extreme scale of the considered phenomenon, corresponding to projections of extreme fields to longer observation periods.

Acknowledgments. We would like to thank Météo-France for providing us the data-set. This work was supported by labEx NUMEV, by the french national program LEFE/INSU and by Montpellier University.

References

[1] Caires, S., de Haan, L., and Smith, R. L. (2011). On the determination of the temporal and spatial evolution of extreme events Technical Report, Deltares. Report 1202120-001-HYE-004 (for Rijkswaterstaat, Centre for Water Management).

[2] Chailan, R., Toulemonde, G., and Bacro, J. (2017). A semiparametric method to simulate bivariate space-time extremes. *The Annals of Applied Statistics* **11(3)**, 1403–1428.

[3] de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory. An Introduction*. Springer Series in Operations Research and Financial Engineering. Springer: New York.

[4] Dombry, C., Engelke, S., and Oesting, M. (2016). Exact simulation of max-stable processes. *Biometrika* **103**, 303–317.

[5] Dombry, C. and Ribatet, M. (2015). Functional regular variations, Pareto processes and peaks over threshold. *Statistics and Its Interface* **8(1)**, 9–17.

[6] Estupina-Borrell, V., Dartus, D., and Ababou, R. (2006). Flash flood modeling with the MARINE hydrological distributed model. *Hydrology and Earth System Sciences* **3**, 3397–3438.

[7] Ferreira, A. and de Haan, L. (2014). The generalized Pareto process; with a view towards application and simulation. *Bernoulli* **20(4)**, 1717–1737.

[8] Thibaud, E. and Opitz, T. (2015). Efficient inference and simulation for elliptical Pareto processes. *Biometrika* **102(4)**, 855–870.

[9] Vinet, F., Boissier, L., and Saint-Martin, C. (2016). Flashflood-related mortality in southern France: first results from a new database. 3rd European Conference on Flood Risk Management (FLOODrisk 2016), Oct 2016, Lyon, France. *E3S Web of Conferences* **7**, article number 06001. DOI: https://doi.org/10.1051/e3sconf/20160706001.

[10] Wang, Y. and Stoev, S. A. (2011). Conditional sampling for spectrally discrete maxstable random fields. *Advances in Applied Probability* **43**, 461–483.

Extra-Parametrized Extreme Value Copula : Extension to a Spatial Framework

J. Carreau[1*] and G. Toulemonde[2]

[1] *HydroSciences Montpellier, Université de Montpellier; julie.carreau@ird.fr*
[2] *Institut Montpellierain Alexander Grothendiek, Université de Montpellier; gwladys.toulemonde@umontpellier.fr*
[*] *Corresponding author*

Abstract. *Extreme-value copulas are justified by the theory of multivariate extremes. However, most high-dimensional copulas are too simplistic for applications. Recently, a class of flexible extreme-value copulas was put forward by combining two extreme-value copulas with a weight parameter in the unit hyper-cube. In a multisite study, the copula dimension being the number of sites, this extra-parametrized approach quickly becomes over-parametrized. In addition, interpolation is not straightforward. The aim of this work is to extend this approach to a spatial framework. By taking the weight parameter as a function of covariates, model complexity is reduced. Moreover, the model is defined at every point of the space and can be interpreted in terms of distances. We focus on the spatial extension based on Gumbel copulas and describe its possible extremal dependence structures. The proposed spatial model is applied on both synthetic and precipitation data in the French Mediterranean.*

Keywords. *Gumbel copula; Spatial extremes; Intense precipitation events.*

1 Regional risk assessment for heavy precipitation

The French Mediterranean is exposed to intense rainfall events called Cevenol events. These regularly cause flooding leading to important material damages and fatalities. Risk assessment is conventionally performed by determining at-site T year return levels - the rainfall intensity level that is expected to be exceeded on average once per T years at a given site. However, for flood risk mitigation, planning is made at a regional scale. To assess risk regionally, regional return levels - levels that can be exceeded anywhere in a given region - can be considered.

To estimate regional return levels, knowledge on the spatial dependence patterns of extreme events is needed. We use a set of 60 rain-gauge stations located in the French Mediterranean for which daily precipitation is available from 1958-2014 (57 years). The exploratory analysis is performed by studying the dependence properties of the annual maxima of daily precipitation at a reference site with respect to annual maxima at all the other sites. For each pair of sites, the bivariate Pickands dependence function $A(\cdot)$ can be estimated [1]. From the estimator of $A(\cdot)$, we compute the extremal dependence coefficient $\chi = 2(1-A(1/2)) \in [0,1]$ which summarizes the strength of the dependence and the extremal asymmetry coefficient $\phi = A'(1/2)/(2(1-A(1/2))) \in [-1,1]$ [3]. When $\phi < 0$ ($\phi > 0$), the reference site tends to take higher (lower) values than the other site, given that both take high values.

In Fig. 1, estimates of the extremal dependence coefficients and the extremal asymmetry coefficients are illustrated for two reference rain-gauge stations. At shorter distances, the χ estimates are about 0.75 indicating asymptotic dependence. In addition, the spatial pattern of extremal dependence - as shown by the white contour level curves of the χ - can be anisotropic and changes with the location of the reference site. The coefficients of extremal asymmetry are meaningful only when the extremal dependence is rather strong, i.e. at short distances. Within the white contour levels, the degree of asymmetry can vary not only with the distance but also with the orientation of the neighbor site.

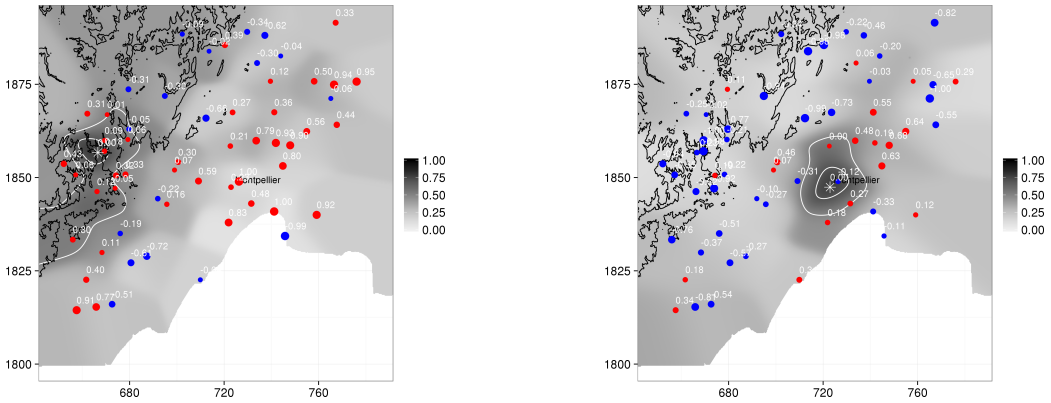


Figure 1: Spatial dependence patterns of annual maxima for two reference sites (white star) in the left and right panels. The grey scale refers to the χ estimates (locally estimated and then smoothed) and two contour levels at $\chi = 0.5$ and $\chi = 0.7$ are shown as white curves. The stations are coloured in red (blue) when the asymmetry is positive (negative) and the estimated ϕ is provided in white above each station.

In order to take into account the aforementioned observations on the spatial pattern of extremal dependence, we propose to model the annual maxima of precipitation with a max-stable process which is an extension to a spatial framework of extra-parametrized multivariate extreme value copula [2]. The proposed max-stable process, once fitted, could be used to simulate long series of fields of annual maxima from which regional return levels can be estimated.

2 Max-stable process with potential asymmetry and anisotropy

The distribution function of the extra-parametrized multivariate Gumbel copula is, for $\boldsymbol{u}$ and $\boldsymbol{a}$ in $[0,1]^d$:

$$C_{\boldsymbol{a}}(\boldsymbol{u}) = C_{\beta_A}(\boldsymbol{u}^{\boldsymbol{a}}) C_{\beta_B}(\boldsymbol{u}^{1-\boldsymbol{a}}), \tag{1}$$

where the vectorial operations are meant componentwise and with $C_\beta(\boldsymbol{u}) = \exp\left\{-\left[\sum_{i=1}^d (-\ln u_i)^\beta\right]^{1/\beta}\right\}$ the multivariate Gumbel copula with parameter $\beta \geq 1$. Since the later is max-stable, it follows that $C_{\boldsymbol{a}}(\cdot)$ is max-stable as well [2]. In order to define the above distribution function for any set of d sites, we let the extra-parameters $\boldsymbol{a} = (a_1, \ldots, a_d)$ be a function of covariates. More precisely, let $\boldsymbol{x}_s$, $1 \leq s \leq d$, be covariate vectors associated to each site. Then the joint distribution at these d sites is given by Eq. (1) with $a_s = a(\boldsymbol{x}_s)$ for a given functional $a(\cdot)$. Because of the symmetries in the distribution function of Eq. (1), we fix $\beta_A \leq \beta_B$ in order to remove redundant sets of parameters.

2.1 Bivariate properties

The Pickands dependence function of the extra-parametrized bivariate Gumbel is

$$A(t) = \underbrace{\left[a_1^{\beta_A}(1-t)^{\beta_A} + a_2^{\beta_A} t^{\beta_A}\right]^{1/\beta_A}}_{A_{\beta_A}(t)} + \underbrace{\left[(1-a_1)^{\beta_B}(1-t)^{\beta_B} + (1-a_2)^{\beta_B} t^{\beta_B}\right]^{1/\beta_B}}_{A_{\beta_B}(t)}. \quad (2)$$

The contribution of the terms $A_{\beta_A}(t)$ and $A_{\beta_B}(t)$ to the asymmetry of the Pickands function is illustrated in the left panel of Fig. 2. The extremal dependence and extremal asymmetry coefficients can be deduced from Eq. (2) :

$$\chi = 2 - [(a_1^{\beta_A} + a_2^{\beta_A})^{1/\beta_A} + ((1-a_1)^{\beta_B} + (1-a_2)^{\beta_B})^{1/\beta_B} \quad (3)$$

$$\phi = \frac{(a_2^{\beta_A} - a_1^{\beta_A})(a_2^{\beta_A} + a_1^{\beta_A})^{1/\beta_A - 1} + ((1-a_2)^{\beta_B} - (1-a_1)^{\beta_B})((1-a_2)^{\beta_B} + (1-a_1)^{\beta_B})^{1/\beta_B - 1}}{2 - [(a_1^{\beta_A} + a_2^{\beta_A})^{1/\beta_A} + ((1-a_1)^{\beta_B} + (1-a_2)^{\beta_B})^{1/\beta_B}]}. \quad (4)$$

The variation of the χ and ϕ coefficients in terms of $0 \leq a_1, a_2 \leq 1$ is illustrated in the middle and right panels of Fig. 2 for $\beta_A = 2$ and $\beta_B = 5$. We note that χ is maximum when $a_1 = a_2$ (along the first diagonal) and increases for decreasing values of the extra-parameter (in the lower left corner). In the limiting case with $a_1 = a_2 = 0$ ($a_1 = a_2 = 1$), the extra-parametrized Gumbel boils down to the Gumbel copula with parameter β_B (β_A). In addition, $\phi = 0$ - the extra-parametrized Gumbel is symmetrical- whenever $a_1 = a_2$.

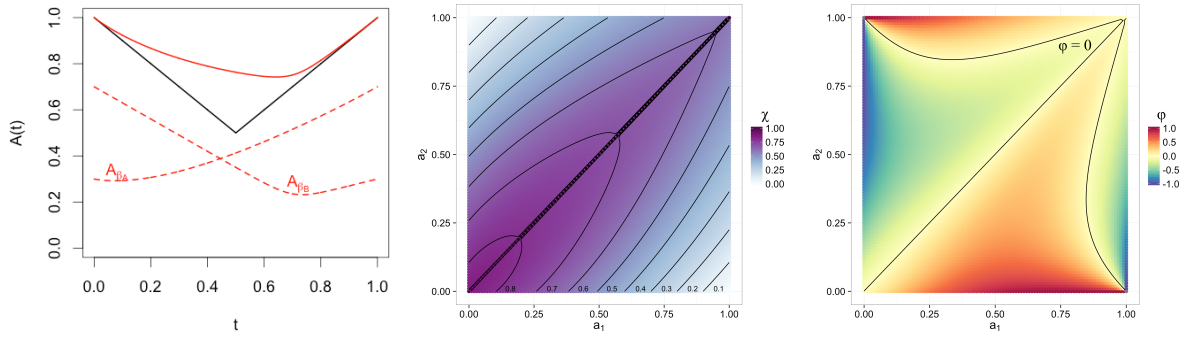


Figure 2: Bivariate properties of the extra-parametrized Gumbel of Eq. (1). Left : Pickands function from Eq. (2) with $\beta_A = 1.5$, $\beta_B = 6$, $a_1 = 0.3$ and $a_2 = 0.7$. Middle and right : χ from Eq. (3) and ϕ from Eq. (4) respectively in terms of $0 \leq a_1, a_2 \leq 1$ with $\beta_A = 2$ and $\beta_B = 5$.

2.2 Inference scheme

We performed a two-step inference scheme. First, univariate generalized extreme-value (GEV) distributions are estimated and serve to transform the annual maxima to Uniform margins. Second, the extra-parametrized Gumbel parameters are estimated by maximizing the pairwise log-likelihood :

$$\sum_{(p_j,p_k)\in\mathcal{P}} \sum_{i=1}^{n} \log\left(c_{\boldsymbol{a}}(z_i(p_j), z_i(p_k))\right) \quad (5)$$

with the first sum running over all the pairs of sites $\mathcal{P}$, the second one over the number of observations n, $z_i(p_j)$ is the i^{th} sample at site p_j with Uniform margins and $c_{\boldsymbol{a}}(\cdot,\cdot)$ the bivariate density associated to

Eq. (1). In the spatial case, the extra-parameters $\boldsymbol{a}$ in Eq. (5) are replaced by their functional form, e.g. linear, and the functional parameters are estimated.

Initial parameters for the multivariate extra-parametrized Gumbel are determined as follows. A dissimilarity matrix is built based on the Spearman ρ estimated for each pair of sites and multidimensional scaling (MDS) is applied to retrieve initial values for the extra-parameters $\boldsymbol{a}$. The initial Gumbel parameters β_A and β_B are determined by estimation on the pairs corresponding to the highest and lowest a_s values respectively, $1 \leq s \leq d$ with d the number of sites. In the spatial case, the parameters of the extra-parameter functional are initialized by performing a regression on the initial values of the extra-parameters obtained by MDS.

3 Synthetic and precipitation data study

The inference scheme is evaluated on synthetic data generated from the model with Uniform margins. The extra-parameters follow a linear mapping while $\beta_A = 2$ and $\beta_B = 5$. A number of sites, from 4 to 60, is randomly sampled from the locations of the rain-gauge stations. The number of observations sampled at each site varies from 20 to 100. We made 1000 replicates for each generative model corresponding to a given number of sites and of observations. The goodness-of-fit is measured in terms of pairwise likelihood on test data and root-mean-square error of the estimated parameters.

Global conclusions from the synthetic data study are the following. The initialization strategy is computationally fast and yields fairly good fit, especially for the smaller training sets. The spatial model outperforms the multivariate model for the larger training sets while being much faster to fit.

For the annual maxima of precipitation at the 60 rain-gauge stations, we first fit the GEV distribution whose three parameters are taken as a function of covariates. The extra-parametrized Gumbel is then fitted to a subset of 8 stations, both in its multivariate and spatial version.

By looking at the bivariate Pickands functions from the fitted models, we conclude that the spatial model is able to reproduce the multivariate model quite well. The shapes of the fitted Pickands display various levels of dependence that can change with the orientation of the stations. In addition, there are several levels of asymmetry. Therefore, these findings stress the need for a max-stable model that can adapt to anisotropy and asymmetry. Further work is required to achieve a satisfactory fit of the extra-parametrized Gumbel jointly at the 60 stations.

References

[1] Marcon, G., Padoan, S.A., Naveau, P., Muliere, P. & Segers, J. (2017). Nonparametric estimation of the Pickands dependence function using Bernstein polynomials, *Journal of Statistical Planning and Inference* **183**, 1–17.

[2] Salvadori, G. & De Michele, C. (2010). Multivariate multiparameter extreme value models and return periods: A copula approach. *Water resources research*, **46**(10).

[3] Semadeni C. & Davison, A. (2015) A Coefficient of Extremal Asymmetry. *Extreme Value Analysis Conference*, Ann Arbor, USA.

Chapter 2

Contributed poster presentations

A dynamic mechanistic species distribution model of wolf recolonization in France

J. Louvrier[1], J. Papaïx[2], C. Duchamp[3] and O. Gimenez[1,*]

[1] *CEFE, CNRS, Univ Montpellier, Univ Paul Valéry Montpellier 3, EPHE, IRD, Fr-Montpellier; julie.louvrier@cefe.cnrs.fr, olivier.gimenez@cefe.cnrs.fr*
[2] *Unité de recherche INRA Biostatistique et processus spatiaux, Fr-Avignon*
[3] *Office National de la Chasse et de la Faune Sauvage, Fr-Gap; julien.papaix@inra.fr*
[*] *Corresponding author*

Abstract. *Species distribution models (SDMs) are a family of statistical tools for ecologists to understand and predict species range. SDMs suffer from several limitations including the difficulty i) to incorporate ecological theory like, e.g., dispersal and ii) to account for imperfect species detectability. Ignoring these issues can lead to bias in inferring species distribution.*
Here, we adopt the theory of ecological diffusion that has recently been introduced in ecological statistics to incorporate in ecological models spatio-temporal processes like dispersal or invasion. As a case study, we focus on wolves (Canis lupus) that have been recolonizing France through dispersal from the Apennines since the last 20 years.
We developed a Bayesian hierarchical model to combine a mathematical formulation of the temporal dynamics of species distribution with data collected in the field. The observation process led to detection/non-detection data that were used to estimate occupancy while accounting for heterogeneity in detection due to variation in abundance. Detection was mainly driven by the sampling effort which we measured as the number of observers per sampling unit. We used differential equations for modelling species diffusion and growth in a fragmented landscape.
We found that our model accurately described the recolonization process of wolves in France. The Bayesian framework was particularly useful to quantify parameter uncertainty in the observation and the ecological processes, and to propagate this uncertainty in the forecasting step.

Keywords. *Bayesian hierarchical model; Imperfect species detection; Mechanistic-statistical modeling; Reaction-diffusion models; Species distribution models.*

---

1 Introduction

Species distribution models (SDMs) are a family of statistical tools for ecologists to understand and predict species range by correlating occurrence data and environmental covariates to produce maps of where species occur and do not occur. SDMs suffer from several limitations including the difficulty i) to incorporate ecological theory like, e.g., dispersal and ii) to account for imperfect species detectability. Ignoring these issues can lead to bias in inferring species distribution.

In what follows, we adopt the theory of ecological diffusion that has recently been introduced in

ecological statistics to incorporate in ecological models spatio-temporal processes like dispersal or invasion [1]. As a case study, we focus on wolves (Canis lupus) that have been recolonizing France through dispersal from the Apennines since the last 20 years. Understanding the mechanisms leading to the expansion of wolves in France can help mitigating conflicts with human activities (attacks on livestock).

2 Model

2.1 Observation process

Let $y_{i,j,t}$ be a random variable that takes value 1 if at least one individual is detected at site $i = 1,\ldots,K$ with spatial location $\mathbf{s}_i$ (the center of site i) within a study area S ($\mathbf{s}_i \in S \subset R^2$) during secondary occasion (or survey) $j = 1,\ldots,J_{i,t}$ in year $t = 1,\ldots,T$, and value 0 otherwise.

Let $n_{i,t}$ be the true abundance. The probability for the species to be detected at site i in year t, $p_{i,t}$, is likely to be influenced by the abundance of the species at a site, $n_{i,t}$. To link the detection/non-detection process to abundance, we used the Royle-Nichols approach [4] which states that if each individual within an occupied area has a detection probability, say r, and there is independence of detections among individuals, then $p_{i,t} = 1-(1-r_{i,t})^{n_{i,t}}$.

We thus have $y_{i,j,t} \sim [y_{i,j,t}|p_{i,t}]$ where $[y_{i,j,t}|p_{i,t}]$ is a probability mass function conditioning occupancy data $y_{i,j,t}$ on the latent, true abundance $n_{i,t}$ through the species-level detection probability $p_{i,t}$. Assuming a binomial observation process, a constant survey effort ($J_{i,t} = J$), and that $r_{i,t}$ and $n_{i,t}$ remains unchanged across the J surveys, we have $\sum_{j=1}^{J} y_{i,j,t} \sim \text{Binomial}(J, p_{i,t})$. The J repeated surveys within each year t are used to estimate the species-level detection probability. Note that if $n_{i,t} = 0$ then $p_{i,t} = 0$ and $y_{i,j,t} = 0\,\forall j$.

Covariates can be incorporated in the individual-level detection probability, $r_{i,t}$. Based on a previous study [2], we used the sampling effort at site i in year t ($\text{Eff}_{i,t}$) and the road density at site i (RoadD_i): $\text{logit}(r_{i,t}) = \beta_0 + \beta_1\,\text{Eff}_{i,t} + \beta_2\,\text{RoadD}_i$.

2.2 State process

We assume that the true abundance is Poisson distributed: $n_{i,t} \sim \text{Poisson}(\lambda(\mathbf{s}_i,t))$ where $\lambda(\mathbf{s}_i,t)$ is the abundance, a spatiotemporal process that describes the dynamics of the number of individuals in site i. In general, we have: $\lambda(\mathbf{s}_i,t) = \int_{B_i} u(x,t)dx$ where $u(x,t)$ is the density at the spatial location x at time t and B_i is the study area in which counts occur. In our case, we can assume the scale at which data were collected coincides with the numerical scale in which we solve $u(x,t)$, precluding the necessity to integrate over B_i.

We used the Partial Differential Equation (PDE) known as ecological diffusion [5] to describe diffusion and growth dynamics. The ecological diffusion PDE in two dimensions with logistic growth is:

$$\frac{\partial u(x,t)}{\partial t} = \Delta(D(x)u(x,t)) + R(x)u(x,t)(1-\frac{u}{K(x)})$$

where Δ is the Laplace 2D diffusion operator (sum of the second derivatives with respect to coordinate). This operator describes uncorrelated random walk movements , with the coefficient $D(x)$ measuring heterogeneous mobility. The term $R(x)$ is the intrinsic growth rate at low density and $K(x)$ is the carrying capacity. In addition we assume reflecting boundary conditions, meaning that the population flows vanish at the boundary of the study site, due to truly reflecting boundaries or symmetric inward and outward fluxes.

2.3 Bayesian inference

To complete the Bayesian specification of the spatio-temporal occupancy-abundance model, we specified non–informative priors for the parameters to be estimated. The approach was implemented in OpenBUGS [3].

3 Results

We found that our model accurately described the dispersal process of wolves in France, and that estimated diffusion of the species was coherent with recolonization (Figure 1).

Figure 1: Estimated abundance in 2007 and 2016.

4 Discussion

The Bayesian framework was particularly useful to quantify parameter uncertainty in the observation and the ecological processes. With the objective to implement an adaptive management strategy for the wolf population, the perspective is to explore the ability of our model to forecast wolf colonization in the future.

Acknowledgments. The authors are sincerely thankful to all the volunteers and wolf regional referees belonging to the Large Carnivore Network for data collection and validation and local investment in the fieldwork. JL is thankful to the GDR Ecologie Statistique, Univ. of Montpellier and ONCFS for grants she received to conduct her work.

References

[1] Hefley, T. J., Hooten, M. B., Russell, R. E., Walsh, D. P. and Powell, J. A. (2017). When mechanism matters: Bayesian forecasting using models of ecological diffusion. *Ecology Letters*, **20**: 640–650.

[2] Louvrier, J., Duchamp, C., Lauret, V., Marboutin, E., Cubaynes, S., Choquet, R., Miquel, C. and Gimenez, O. (2017). Mapping and explaining wolf recolonization in France using dynamic occupancy models and opportunistic data. *Ecography*.

[3] Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009). The BUGS project: Evolution, critique and future directions (with discussion), *Statistics in Medicine* **28**: 3049–3082.

[4] Royle, J. A. and Nichols, J. D. (2003). Estimating abundance from repeated presence-absence data or point counts. *Ecology*, **84**: 777–790.

[5] Williams, P. J., Hooten, M. B., Womble, J. N., Esslinger, G. G., Bower, M. R. and Hefley, T. J. (2017). An integrated data model to estimate spatiotemporal occupancy, abundance, and colonization dynamics. *Ecology*, **98**: 328–336.

Investigating the relationship between fluid injection and triggered seismicity in southern Italy

L. Telesca[1,*] and T.A. Stabile[1]

[1] *National Research Council, Institute of Methodologies for Environmental Analysis, C.da S. Loja, 85050 Tito (PZ), Italy; luciano.telesca@imaa.cnr.it*
[*] *Corresponding author*

Abstract. *In this study the relationship between injected wastewater produced during oil exploitation and triggered seismicity is analysed by using the Schuster's spectrum of the earthquake point process and the singular spectrum analysis (SSA) of the volume and pressure of the wastewater. Our findings show that most of the periodicities identified by Schuster's spectrum of earthquakes coincide with those identified by the SSA of volume and pressure of the injected wastewater, indicating that these are able to excite or induce oscillatory fluctuations in the seismic rate, and so, to generate earthquakes.*

Keywords. *Point processes; Earthquakes; Schuster's test; Singular spectrum analysis.*

---

1 Introduction

Seismicity induced by the human activity has been documented since the 1920s. In recent years, injection-induced earthquakes raised great concern. For instance, in the USA the largest injection-induced earthquake with magnitude 5.7 occurred in Oklahoma, being the largest induced earthquake in the world. Although efforts have been made to comprehend the role of fluids in the triggering processes of human-induced seismicity, indentifying whether an earthquake is triggered by energy technologies is still challenging. In this study, we focus on the seismically active area of High Agri valley (southern Italy), which hosts the biggest onshore oil field in west Europe, with an average production of about 3.6×10^9 kg of oil per year. On 2 June 2006 the wastewater produced during the oil exploitation started to be pumped back into a 4km deep injection well; and after only 4 days the seismic activity around increased significantly. To better investigate the possible link between earthquakes and fluid-injection activities, we applied the Schuster's test to the earthquakes and the singular spectrum analysis (SSA) to the fluid-injection variables (pressure and volume). Our aim is to identify common periodic behaviours in both processes, which could suggest that oscillatory fluctuations in seismicity might be triggered by fluid-injection periodic forcings. The investigation of periodic behavior is a challenging topic in seismology, since periodic earthquake occurrences may reflect links with semidiurnal to multiyear tides, seasonal hydrological loads or 14 month pole tide forcing [1]. So far, no studies have been perfomed in investigating periodicities in fluid-injection triggered seismicity.

2 Methods

Considering a series of seismic events, where t_k represents the occurrence time of the k-th event of the series, its associated phase θ_k is given by

$$\theta_k = 2\pi\frac{t_k}{T}$$

where T is the periodicity to test. A two-dimensional walk can be constructed from the series of earthquake occurrence times, whose associated phases drive the successive directions of a unit-length step. Indicating with D the distance between the start and end points of this walk, the probability p that a distance greater than or equal to D can be reached by a uniformly random two-dimensional walk, is the probability of the null hypothesis that the distribution of the time occurrences arises from a uniform seismicity rate [4]:

$$p = e^{\frac{-D^2}{N}}$$

where N is the number of events. The probability p is called Schuster p-value, and the lower its value, the higher the probability that the occurrence times are governed by the periodicity T. The periodicity T can be detected above the 95% confidence level if the corresponding Schuster's p-value is lower than $0.05 \times T/t$, where t is the period of the sequence [2].

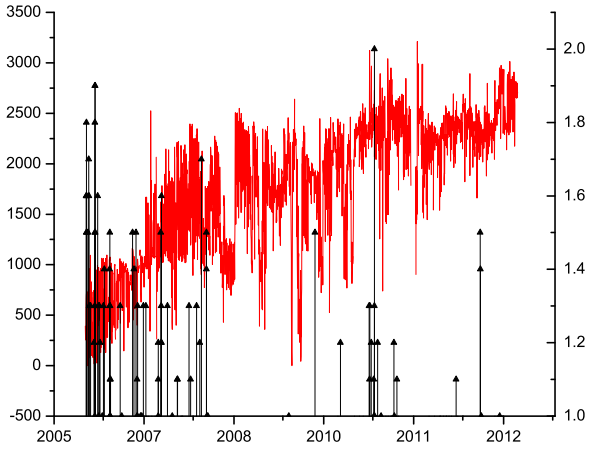


Figure 1: Seismicity (black vertical arrows) and daily variation of volume (red) during the investigation period.

The Singular Spectrum Analysis (SSA) [5] is a well know decompositional method of irregular time series, aiming at decomposing a signal into a certain number of independent components (trend, oscillatory components and structureless noise). Given the series $y_i, i = 1, \cdots, N$, where N is the length of the series, the SSA is based on the calculation of its Toeplitz lagged correlation matrix with lag W that is the number of independent components in which the series has to be decomposed. The eigenvalues λ_k of the Toeplitz matrix are sorted in decreasing order for k varying from 1 to W, and correspondingly the eigenvectors E_{jk} are determined, for j varying from 1 to W. The reconstructed components r_{ik} of the y_i are then calculated

$$r_{ik} = \frac{1}{W}\sum_{j=1}^{W} a_{i-j,k}E_{jk}, \quad W \le i \le N - W + 1$$

where a_{ik} is the k-th principal component given by

$$a_{ik} = \sum_{j=1}^{W} y_{i+j} E_{jk}, \quad 0 \leq i \leq N - W$$

The eigenvalue λ_k represents the fraction of the total variance of y_i hold by the reconstructed component r_{ik}. Typically, most of the total variance of the series is contained in the first reconstructed components that usually represent the slowly varying trend and the quasi-oscillatory components, while the remaining ones represents structureless noise.

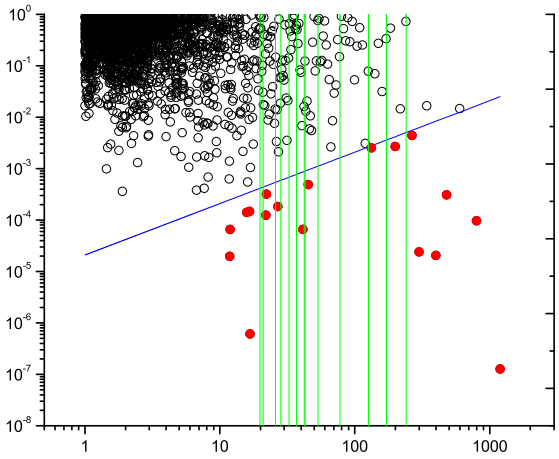


Figure 2: p-value of the tested period above (black circles) and below (red circles) the 95% confidence level (blue line). The green vertical lines indicate the periods of the volume of the injected wastewater.

3 Results

We applied the Schutser's spectrum to the point process of earthquakes with magnitude $ML \geq 1.1$ testing the periods between 1 and 1196 days, and the SSA to volume and pressure of the injected fluid. Figure 1 shows the earthquake series and the time variation of the volume. Figure 2 shows the Schuster's p-value (black hollow circles) versus the period tested for earthquakes; the blue line represents the 95% confidence detection level (a periodicity is significant at 95% confidence if its p-value lies below the 95% confidence detection line (red circles)). Few periods are below the 95% confidence detection level and may be associated to periodicities in earthquakes. We applied the SSA to volume V and pressure P with lag $W = 169$ chosen on the basis of the criterion stated in [3]. Only few significant components are kept by applying the minimum description length (MDL) criterion [6]:

$$MDL(k) = -log\left(\frac{\prod_{i=k+1}^{W} \lambda_i^{\frac{1}{W-k}}}{\frac{1}{W-k}\sum_{i=k+1}^{W} \lambda_i}\right)$$

where k is the order of the eigenvalue λ_k, W is the number of eigenvalues and N is the length of the original series. The minimum of MDL was at $k = 19$ for V and $k = 21$ for P. For each significant com-

ponent we calculated the periodogram and identified the period corresponding to its maximum. Figure 2 shows the periods of the volume, while Figure 3 show those of the pressure. We can see that there is a good agreement between most of the periods of the earthquake sequence (red circles) and those of the volume and pressure (green vertical lines). We can see that most of the periods identified by Schuster's test on the earthquakes coincide or are very close to those identified by the SSA on V and P, indicating that the periodicities contained in these variables are able to excite or induce oscillatory fluctuations in the seismic rate and so to generate earthquakes.

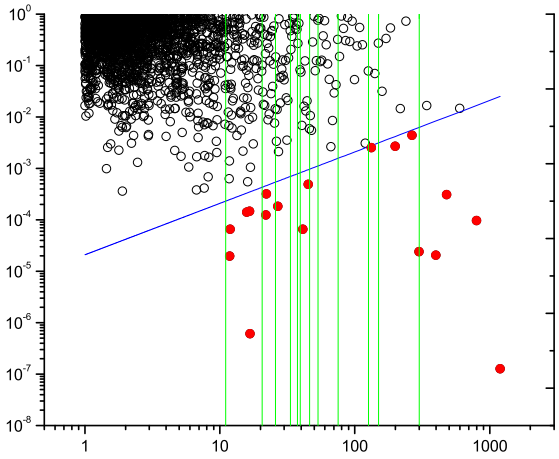


Figure 3: p-value of the tested period above (black circles) and below (red circles) the 95% confidence level (blue line). The green vertical lines indicate the periods of the pressure of the injected wastewater.

4 Conclusions

In this study we applied two robust statistical tools to asses the significant relationship between injected wastewater and triggered seismicity. The obtained results could contribute to a better comprehension of seismicity generated by human technologies.

References

[1] Dutilleul, P., Johnson, C. W., Burgmann, R., Wan, Y. and Shen, Z.-K (2015). Multifrequential periodogram analysis of earthquake occurrence: An alternative approach to the Schuster spectrum, with two examples in central California. *Journal of Geophysical Research* **120** 8494–8515.

[2] Ader, T. J. and Avouac, J.-P. (2013). Detecting periodicities and declustering in earthquake catalogs using the Schuster spectrum, application to Himalayan seismicity. *Earth and Planetary Science Letters* **377-378** 97–105.

[3] Khan, M. A. R. and Poskitt, D. S. (2010). Description length based signal detection in singular spectrum analysis. *Monash Econometrics and Business Statistics Working Papers* **13/10**

[4] Schuster, A. (1897). On lunar and solar periodicities of earthquakes. *Proceedings of the Royal Society of London* **61** 455–465.

[5] Vautard, R, Yiou, P., Ghil, M. (1992). Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D* **58** 95–126.

[6] Wax, M. and Kailath, T. (1985). Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **33** 387–392.

Varying coefficient models for areal data

M. Franco-Villoria[1,*], M. Ventrucci[2] and H. Rue[3]

[1] *University of Torino; maria.francovilloria@unito.it,*
[2] *University of Bologna; massimo.ventrucci@unibo.it*
[3] *King Abdullah University of Science and Technology; haavard.rue@kaust.edu.sa*
[*] *Corresponding author*

Abstract. *We discuss the use of penalized complexity priors for spatially varying coefficient models, introducing a natural base model choice that corresponds to a constant coefficient (no variation in space). Preliminary results on the use of these priors in a case study on air pollution and hospital admissions in Turin (Italy) are presented.*

Keywords. *Varying Coefficient models; PC prior; ICAR*

1 Introduction

Varying coefficient models (VCMs) [2] are useful in the presence of a variable that *modifies* the effect of a covariate of interest on the response. Consider the simple case where there are n observational units indexed by $i = 1,\dots,n$ and one covariate x_i whose effect on the response y_i depends on another variable z_i. Assuming y_i belonging to the exponential family, the linear predictor of a generalized VCM is

$$\eta_i = \alpha + f(z_i)x_i \qquad i = 1,\dots,n. \tag{1}$$

We follow a Bayesian hierarchical framework where the varying coefficient $f(z_i)$ in Eq. (1) is described by a vector of random effects $\boldsymbol{\theta} = (\theta_1,\dots,\theta_n)^\intercal$ distributed at prior as a Gaussian Markov Random Field (GMRF) [4] with sparse precision $\boldsymbol{Q}(\boldsymbol{\tau}) = \tau\boldsymbol{R}$. For areal data, the index $i = 1,\dots,n$ indicates each of the non overlapping regions in a lattice. The spatially varying coefficient $\boldsymbol{\theta} = (\theta_1,...,\theta_n)^\intercal$ follows an Intrinsic Conditional Autoregressive (ICAR) model [1]:

$$\theta_i|\boldsymbol{\theta}_{-i},\tau \sim N\left(\sum_{j:i\sim j}\frac{\theta_j}{n_i},(n_i\tau)^{-1}\right)$$

where $i \sim j$ denotes neighbouring regions (sharing a common border) and n_i denotes the number of neighbours of region i. The joint distribution for $\boldsymbol{\theta}$ is given by $\boldsymbol{\theta}|\tau \sim N\left(0,(\tau\boldsymbol{R})^{-1}\right)$ where the structure matrix $\boldsymbol{R}$ is a singular matrix with entries:

$$R_{i,j} = \begin{cases} n_i & i = j \\ -1 & i \sim j. \end{cases}$$

2 Penalized complexity (PC) priors for spatially varying coefficients

A useful parametrization for the varying coefficient is $\theta_i = \beta + \delta_i$, where δ_i indicates deviation from a constant slope β at value z_i. When $\delta_i = 0\ \forall i$, the varying coefficient $f(z)$ is constant over z and model (1) becomes a simple linear regression model (*base model*), while the VCM can be seen as a flexible extension of it; this flexibility is regulated by the precision parameter τ. Simpson et al. (2017) [5] recently introduced a new framework for building priors that avoid overfitting denoted as *Penalized Complexity (PC) priors*. PC priors are computed based on specific principles in which a model component is seen as a flexible parametrization of a base model. Model complexity, defined in terms of distance from the base model, is penalized so that the base model is favoured unless the data support a more flexible one.

If we consider $f(z)$ in terms of the vector of random effects $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\mathsf{T}$ introduced in Section 1, the base model can be obtained setting the hyper-parameters $\boldsymbol{\tau}$ to a particular value. When $\tau^{-1} = 0$ we have $f(z_i) = \beta$, which implies the linear regression model, $\eta_i = \alpha + \beta x_i$. For $\tau^{-1} > 0$, $f(z_i)$ incorporates higher degree of complexity w.r.t. the constant slope, leading to the flexible VCM. The PC prior is an exponential distribution on the distance scale (measured using the Kullback-Leibler divergence [3]). A change of variable gives the PC prior in the scale of the precision. For a generic Gaussian Random effect conditional on τ, the PC prior for τ is the Gumbel type 2 with density,

$$\pi(\tau) = 0.5\lambda\tau^{-1.5}\exp(-\lambda\tau^{-0.5}); \tag{2}$$

for more details see [5]. The parameter λ in Eq. (2) can be selected through a user-defined scaling approach, by setting U and α such that $Pr(1/\sqrt{\tau} > U) = a$, which yields $\lambda = -\log(a)/U$ [5] .

3 Application: PM_{10} and hospital admissions in Torino, Italy

Data on daily hospital admission due to respiratory causes are available from hospital discharge registers for the 315 municipalities in the province of Torino, Italy in 2004. On the other hand, daily particular matter PM_{10} ($\mu g/m^3$) data and average temperature (Kelvin degrees) are available at municipality level. We consider a Poisson regression model that includes spatially structured random effects to estimate the effect of PM_{10} on hospitalization risk; the PM_{10} effect is allowed to vary across municipalities and the PC prior in Eq. (2) is used for the precision parameter of the varying coefficient. Preliminary results (not shown here) suggest that the posterior relative risk changes across municipalities.

References

[1] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society. Series B*, **36(2)**:192–225.

[2] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B*, **55(4)**:757–796.

[3] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**:79–86.

[4] Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman and Hall/CRC.

[5] Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, **32(1)**:1–28.

A space-time branching process with covariates

Giada Adelfio[1,2,*], Marcello Chiodi[1,2]

[1] *Dipartimento di scienze Economiche Aziendali e Statistiche, University of Palermo;*
[2] *Istituto Nazionale di Geofisica e Vulcanologia (INGV), Palermo;*
giada.adelfio@unipa.it, marcello.chiodi@unipa.it
[*] *Corresponding author*

Abstract. *The paper proposes a stochastic process that improves the assessment of seismic events in space and time, considering a contagion model (branching process) within a regression-like framework. The proposed approach develops the Forward Likelihood for prediction (FLP) method including covariates in the epidemic component.*

Keywords. *Space-time Point Process; FLP; covariates; ETAS model*

---

1 Introduction

Contagious phenomena are well described in space and time by self-exciting point processes, where the conditional intensity function is obtained as the sum of the long-term variation component (called endemic) and the short-term variation one (named epidemic). This kind of models have been widely used in the literature: infectious disease [10], crime [11], quakes [12], [1]. To model earthquake activity in space and time accounting both for the endemic (background activity) and epidemic (aftershocks) effect, the Epidemic-Type Aftershock Sequences (ETAS) model is used. It describes events starting from their space-time coordinates (and magnitude as mark) and incorporates seismological laws in a mechanistic approach (e.g. the Omori law) as a natural one in the context of earthquake data. In this paper, we aim at providing an improved computational framework for further theoretical and empirical developments for studying and describing epidemic phenomena, where there is a contagious effect of the previous history, in space an time, and of specific covariates. In particular, we suggest the use of a branching-type model for earthquake description (the ETAS model) in a regression-oriented version modelling, accounting also for external covariates, for explaining some of the overall variability of the studied phenomenon and reducing the unpredictable variability. We provide developments of the Forward Likelihood for prediction (FLP) method [5] for estimating the ETAS model components, introducing covariates for the epidemic part, for a more realistic description of observed patterns.

2 Branching point processes and ETAS model with covariates

Branching processes are used to model reproduction phenomena. These models have been recently considered for the description of different applicative fields: biology [4], demography [9], epidemiology

[3]. Any analytic space-time point process defined in $[0,T] \times W \subset \mathbf{R}^2, T > 0$ is uniquely characterized by its associated *conditional intensity function* (CIF) [8]:

$$\lambda(t,\mathbf{s}|H_t) = \lim_{\Delta t,\Delta \mathbf{s}\to 0} \frac{\mathrm{E}\left[N([t,t+\Delta t],[\mathbf{s},\mathbf{s}+\Delta \mathbf{s}]|H_t)\right]}{\Delta t \Delta \mathbf{s}}$$

where H_t is the space-time occurrence history of the process up to time t, $\Delta t, \Delta \mathbf{s}$ are time and space increments, $\mathrm{E}\left[N([t,t+\Delta t],[\mathbf{s},\mathbf{s}+\Delta \mathbf{s}]|H_t)\right]$ is the history-dependent expected number of events occurring in the volume $\{[t,t+\Delta t) \times [\mathbf{s},\mathbf{s}+\Delta \mathbf{s}]\}$. Generally, intensities $\lambda(\cdot)$ depend on some unknown parameter $\boldsymbol{\theta}$, so that we have $\lambda(\cdot,\boldsymbol{\theta})$. The CIF represents the instantaneous rate or hazard for events at time t and location $\mathbf{s}$ given all the observations up to time t, conditioning on the random past history of the process. In general, the conditional intensity function of the branching model is defined as the sum of a term describing the large-time scale variation (spontaneous activity or background) and one relative to the small-time scale variation due to the interaction with the events in the past (induced activity or offsprings):

$$\lambda_{\boldsymbol{\theta}}(t,\mathbf{s}|H_t) = \mu f(\mathbf{s}) + \tau_{\phi}(t,\mathbf{s}) \tag{1}$$

with H_t the past history of the process, $\boldsymbol{\theta} = (\phi,\mu)'$, the vector of parameters of the induced intensity (ϕ) together with the parameter of the background general intensity (μ), $f(\mathbf{s})$ the space density, and $\tau_{\phi}(t,\mathbf{s})$ the induced intensity (or self-exciting component), given by:

$$\tau_{\phi}(t,\mathbf{s}) = \sum_{t_j<t} \nu_{\phi}(t-t_j,\mathbf{s}-\mathbf{s}_j).$$

The self-exciting component of the model essentially provides a description of the intensity at a space-time location $(t,\mathbf{s})$ caused by each previous event. In such models, we have to simultaneously estimate the different components of the intensity function (large-time scale and small-time scale). If the large-time scale component $\mu f(\mathbf{s})$ in (1) is known, the parameters ϕ can be usually estimated by Maximum Likelihood method. In applications, the large-time scale component $\mu f(\mathbf{s})$ is usually estimated trough nonparametric techniques, like kernel estimators.

In seismological context, the branching process ETAS model has been introduced [12]. Starting from model (1), the ETAS conditional intensity function can be written as follows:

$$\lambda_{\boldsymbol{\theta}}(t,\mathbf{s}|H_t) = \mu f(\mathbf{s}) + \sum_{t_j<t} g(t-t_j|m_j)\ell(\mathbf{s}-\mathbf{s}_j|m_j) \tag{2}$$

with m_j the magnitude of the j-th event, $g(\cdot)$ the Omori law for occurrence density of aftershocks in time and $\ell(\cdot)$ the spatial distribution, conditioned to magnitude of the generating event. Since the criticality of the simultaneous estimation of the background intensity and the triggered intensity components of a Epidemic type model, the FLP approach was developed (see [5], [7]). It is a nonparametric estimation procedure, used for the large time scale component, based on the subsequent increments of log-likelihood obtained adding an observation one at a time, to account for the information of the observations until t_k on the next one. That provides a simultaneous estimation of the two parametric components of a branching-type model, alternating the standard likelihood method, to estimate the parameters, with the FLP approach, to estimate the nonparametric part. Given the lack of specific open-source tools, the package etasFLP [7] [6] provides tools to implement this mixed approach for a wide class of ETAS models for the description of seismic events, developed in the R environment.

In this paper, we propose an additive-multiplicative model for the conditional intensity function of a space-time point process, incorporating a forward predictive likelihood estimation approach for semi-parametric intensity function. Starting from the definition provided in eq. (1), we propose to modify

the offspring component, accounting for a vector of covariates. As proposed by [10] in a context of infection occurrences, we incorporate the space-time phenomenological laws of the triggering part of the ETAS model with the effects of covariates. This triggering function is factorized into separate effects of external information, time and relative location, such that:

$$\lambda_{\boldsymbol{\theta}}(t,\mathbf{s}|H_t) = \mu f(\mathbf{s}) + \sum_{t_j<t} exp(\eta_j)\tilde{g}(t-t_j|m_j)\tilde{\ell}(\mathbf{s}-\mathbf{s}_j|m_j) \tag{3}$$

where $(t_j, \boldsymbol{s}_j)$ is the time and location of individual occurrence j, $\eta_j = \boldsymbol{\beta}'\boldsymbol{z}_j$ is a linear predictor based on the vector of unpredictable variables $\boldsymbol{z}_j$ of each event, and $\tilde{g}$ and $\tilde{\ell}$ are defined as in eq (2), accordingly modified. In the seismic context, the proposed approach would provide a more general formalism for the earthquake occurrence in space and time. Indeed, the main idea is that the effect on the future activity does not depend only on the closeness of the previous events, but also on specific characteristics of the main event, like magnitude, as usual, and also further information, such as geological features.

3 Application to the Italian earthquakes and comments

We report some of the results of the proposed ETAS-FLP approach with covariates, starting from the Italian catalogue of the space-time Italian seismicity, from May 5th, 2012 to May 7th, 2016, with 2.5 as the threshold magnitude (i.e. the lower bound for which earthquakes with higher values of magnitude are surely recorded in the catalogue). The catalogue reports the usual hypocentral coordinates (longitude, latitude, depth, time) together with the magnitude of the events, and also some additional information, such as: the hypocentral uncertainty, the distance from the nearest station (for shallow earthquakes, this distance should be sufficiently small), a measure of the quality of the location (named rms), the number of stations that recorded the event (this number is heavily influenced by the magnitude of the event and strongly influences the accuracy of the location) and the distance from the nearest fault (i.e. the identified earthquakes sources in that area). Estimating the ETAS model as in eq. (2) by using the FLP approach and accounting for the epicentral coordinates (longitude, latitude and time) and the magnitude of the inducing event, results (not here reported) are not completely satisfying, suggesting, as usual, some lack of fitting mostly due to the triggered component. However, adding also the available covariates, that is considering the model in eq. (3), the best estimated one includes, together with the magnitude, also the depth, the distance from the nearest station and the distance from the nearest fault. In particular, the last two covariates have both a negative effect on the space-time reproducing activity. Diagnostic results suggest a satisfying fitting, as shown in the Figure 3 (see [2] for the residual description).

The reported results, tough partial and provisional, confirm our intuition reported in previous studies (e.g. [2]). Indeed, the need of a more flexible model for the space-time triggered component of the ETAS model is often revealed, although the background seismicity is well described by the FLP estimated intensity. In our opinion, considering external information (such as geological information related to faults distribution) for the description of spatio-temporal earthquakes is a innovative and promising perspective of study, even relevant in different fields of research.

Acknowledgments. This paper has been supported by the national grant of the Italian Ministry of Education University and Research (MIUR) for the PRIN-2015 program “Prot. 20157PRZC4. PI: G.Adelfio".

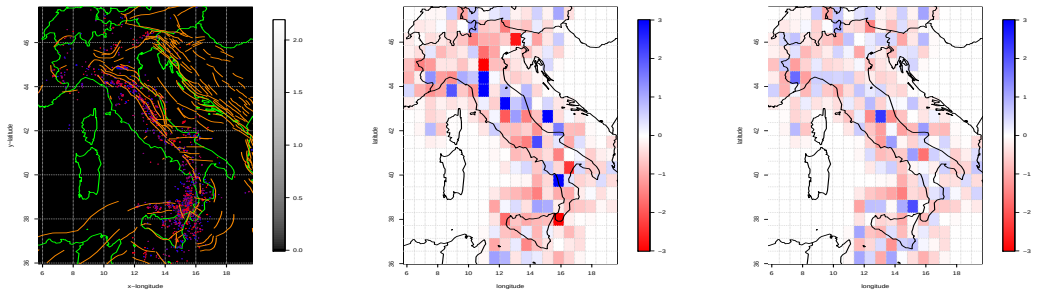


Figure 1: Output for the estimated ETAS-FLP model with covariates: estimated total intensity together with the observed points (old in blue and recent in red) and available line-faults (on the left); space residuals for the model (in the middle), space residuals for the background (on the right).

References

[1] G. Adelfio and M Chiodi, *Alternated estimation in semi-parametric space-time branching-type point processes with application to seismic catalogs*, Stochastic Environmental Research and Risk Assessment **29** (2015), 443–450.

[2] G. Adelfio and M. Chiodi, *Flp estimation of semi-parametric models for space-time point processes and diagnostic tools*, Spatial Statistics **14** (2015), 119–132.

[3] E. Balderama, F.P. Schoenberg, E. Murray, and P.W. Rundel, *Application of branching point process models to the study of invasive red banana plants in costa rica*, JASA **107** (2012), 467–476.

[4] G. Caron-Lormier, J.P. Masson, N. Menard, and J.S. Pierre, *A branching process, its application in biology: Influence of demographic parameters on the social structure in mammal groups*, Journal of Theoretical Biology **238** (2006), 564–574.

[5] M. Chiodi and G. Adelfio, *Forward likelihood-based predictive approach for space-time processes*, Environmetrics **22** (2011), 749–757.

[6] M. Chiodi and G. Adelfio, *etasflp: Estimation of an etas model. mixed flp (forward likelihood predictive) and ml estimation of non-parametric and parametric components of the etas model for earthquake description*, R package version 1.4.1. (2017).

[7] M. Chiodi and G. Adelfio, *Mixed non-parametric and parametric estimation techniques in R package etas-FLP for earthquakes' description*, Journal of Statistical Software **76** (2017), no. 3, 1–28.

[8] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes*, second ed., New York: Springer-Verlag, 2003.

[9] R. A. Johnson and J. R. Taylor, *Preservation of some life length classes for age distributions associated with age-dependent branching processes*, Statistics and Probability Letters **78** (2008), 2981–2987.

[10] S. Meyer, L. Held, and M. Hohle, *Spatio-temporal analysis of epidemic phenomena using the r package surveillance*, Journal of Statistical Software **77** (2012), no. 11.

[11] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, *Self-exciting point process modeling of crime*, Journal of the American Statistical Association **106** (2011), no. 493, 100–108.

[12] Y. Ogata, *Space-time point-process models for earthquake occurrences*, Annals of the Institute of Statistical Mathematics **50** (1998), no. 2, 379–402.

Modelling the environmental risk on the evolution wildfires using random spread process including covariates

Carlos Díaz-Avalos[1], Pablo Juan[2]*, Laura Serra-Saurina[3] and Pau Aragó[2]

[1] *UNAM, Mexico; zhangkalo@gmail.com*
[2] *Department of Mathematics of UJI, Castellón, Spain; juan@uji.es; arago@uji.es*
[3] *Center for Research in Occupational Health (CiSAL), University Pompeu Fabra, Barcelona, Spain; laura.serra-saurina@upf.edu*
[*] *Corresponding author*

Abstract. *The objective of the present study is first, the last developments and validation results of modeling the evolution of wildfires using random spread process. Second, a sensitivity analysis was conducted to identify the most influential covariates for controlling fire propagation. The model combines the features of a network model with those of a quasi-physical model of the interaction between burning and non-burning cells, which strongly depends on covariates. The models applied to different wildfires in Spain, including the different temporal states. In the same way, the possible predictions for compare the experiments in terms of rate of spread, area and shape of the burn and it is studied the environmental risk during the fire propagation. Finally, the sensitivity of the model outcomes to input parameters is modeled. In this work, the idea of random set modeling of fire spread is developed, including the covariates. Some facts indicating the stochastic nature of fire spread are reviewed. A brief survey of deterministic and stochastic models of spread and a description of random set models based on a Markov process called random spread process (RSP) is developed too.*

Keywords. *Covariates; Environmental Risk; Random Spread Process; Wildfires.*

---

1 Introduction

Landscape patterns are determined by the frequency, intensity and extent of disturbances. Wildfires play an important role on this regard, because they have the potential to affect severely the forest dynamics and to produce geohydrological changes. At landscape scale, ignition location and burned area by forest fires are the result of a complex interaction between climatic factors, topography and land (Moreira et al., 2011). Studying the rate of spread and the final shape of the burned area is an important task as this gives insight on whether or not a given wildfire will pose a potential hazard to human life an property as well as to where its contribution to green house gasses and to particulate matter in the atmosphere is more likely to have an impact.

To model fire spread, we consider a study area $\mathcal{W}$ divided in $N_{\mathcal{G}}$ pixels or sites. A question of interest is: given that a fire has ignited at a given site, which factors influence its spread direction and its final shape?

Most fire spread models consider the physics of the burning process and include approaches based on

the physics of fire and in raster spread processes (Bin Sullivan, 2010; Drissi, 2015; Muzy et al., 2008 and Vorob'ov, 1996). Physicists use concepts and mathematical models based on partial differential equations and ordinary differential equations to describe fire spread process (Pastor et al. 2003). Parameters of those equations however do not correspond directly to the parameters intuitively identified as influencing the fire spread, for example, slope, wind speed, biomass, and relative humidity of the vegetation at the fire front. The inclusion of covariates in fire spread models is likely to improve model quality (Aragó et al. 2016; Díaz-Avalos, et al. 2016).

2 Data sets and methods

The data for this project came from two wildfires: The wildfire of Artana, Castellon (Spain) and the Carcaixent wildfire in Alicante. Figure 1 shows the shape of the burned area for the Artana wildfire at 9 consecutive days (upper panel), the shapes and sizes overlaid (lower panel) . Figure 2 shows the shape of the burned area for the wildfire in Carcaixent, Alicante (Spain) at five consecutive days (upper panel) and the shapes and sizes for those days overlaid (lower panel).

For these data sets we consider a model based on a raster approach, where for a given time t, every pixel is in one of three sets: burned (B), burning at the fire front (F) or within the range of the fire front, either because the pixel is in the neighborhood set of a burning pixel, or because it can be reached by sparks for example (N). Note that for $t = t+1$ some pixels in the complement of $B \cup F \cup N$ may join F or N.

To describe the features of wildfire spread, in this study we use a Markovian random field process where at a given time t, each cell can be in any of three states: unignited cells (U), either neighboring the fire front that can be reached by sparks transported by physical forces; burning cells (F); burnt out cells (B), this is, $\mathcal{G} = U \cup F \cup B$. Conditional on F, the probability that unignited cells ignite is a function the local conditions through covariate values of the form

$$log\left[\frac{p(x_{t+1} \in F | x_t \in U)}{1 - p(x_{t+1} \in F | x_t \in U)}\right] = Z\beta + U(x_t) \tag{1}$$

where Z is a matrix with covariates related to the risk of fire ignition and $U(\cdot)$ is a spatial term that depends on the pixels in F within the neighborhood of x_t. The model was fitted using a bayesian approach. Model results ad future research work are discussed.

3 Acknowledgments

We would like to thank the Environment Department of the Government of Valencia for the access to the digital map databases.

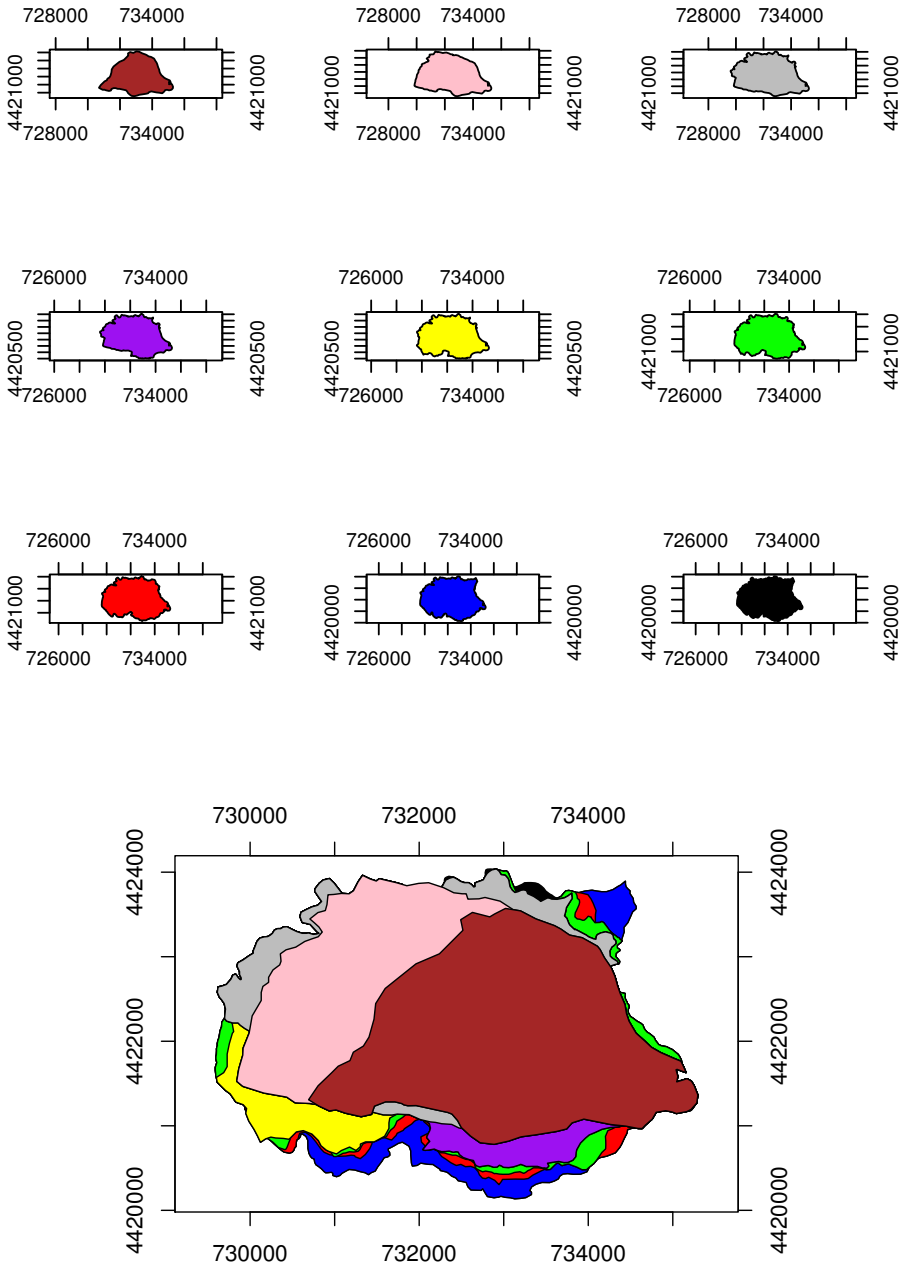


Figure 1: Time evolution of a wildfire in Artana, Castellon, Spain, 9 instants (top) and all step-time in the same Figure (bottom)

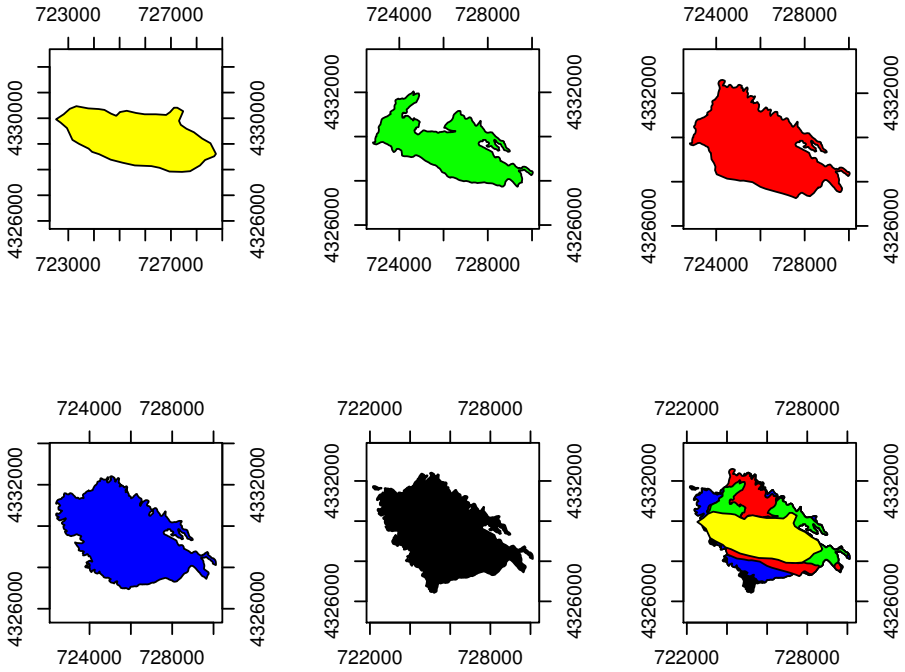


Figure 2: Wildfire in Carcaixent, Alicante, Spain, in 5 instants and the shape of the total burned area.

References

[1] Aragó, P., Juan, P., Díaz-Avalos, C., Salvador, P. (2016). Spatial point process modeling applied to the assessment of risk factors associated to forest wildfires incidence in Castellón, Spain. *European Journal of forest reserach.*

[2] Bin Suliman, M. D. H. and Serra, J. and Mahmud, M. (2010). Prediction and simulation of Malaysian forest fires by random spread. International Journal of Remote Sensing. **31 (22)**. p. 6015-6032.

[3] Díaz-Avalos, C., Juan, P., Serra-Saurina, L. (2016). Modeling fire size of wildfires in Castellon (Spain), using spatiotemporal marked point processes. *Forest Ecology and Management*, **381**, 360-369.

[4] Drissi, Mohamed, (2015). Modeling the spreading of large-scale wildland fires. *Wildland fires conference* Missoula, MT. Proc. RMRS-P-73. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. p. 278-285.

[5] Moreira, F., Viedma, O., Arianoutsou, M., Curt, T. and Koutsias, N. Landscape - wildfire interactions in southern Europe: Implications for landscape management. *Journal of Environmental Management*, Elsevier, 2011, 92, p. 2389 - p. 2402.

[6] Muzy, A.,J. Nutaro, B.P. Zeigler, and P. Coquillard, (2008). Modeling and simulation of fire spreading through the activity tracking paradigm. *Ecol. Model.*, **219(1-2)**. p. 212-225.

[7] Pastor E., Zàrate L., Planas E., Arnaldos J. (2003). Mathematical models and calculation systems for the study of wildland fire behaviour. *Prog. Ener. Combust. Sci.* **29**, 139-153.

[8] Vorob'ov, O. Yu. , (1996). Random set models of fire spread *Fire Technology*, Springer Netherlands, **32**, Number 2.

Bayesian space-time modeling of multivariate marine litter abundance

C. Calculli [1,*], A. Pollice[1], L. Sion [2] and P. Maiorano [2]

[1] *Department of Economics and Finance, University of Bari Aldo Moro, Bari, Italy; crescenza.calculli@uniba.it, alessio.pollice@uniba.it*
[2] *Department of Biology, University of Bari Aldo Moro, Bari, Italy; porzia.maiorano@uniba.it, letizia.sion@uniba.it*
[*] *Corresponding author*

Abstract. *In recent years, marine litter has become a recognized global ecological concern, although its distribution and influence on deep-sea habitats are still not well-known. This study focuses on the analysis of abundance data for litter categories, collected during trawl surveys regularly conducted at local scale, in the Central Mediterranean. Multivariate abundance data have space-time structure and come with additional environmental continuous covariates, such as distances to the coastline or to the nearest harbor. Here marine litter data are modeled in order to estimate the effects affecting the dynamics of litter assemblages at different spatio/temporal scales. We propose a correlated response model with latent variables, that proves to be particularly suitable to infer potential environmental covariates while controlling for correlation between litter categories and providing a method for residual ordination. MCMC estimation is implemented within the Bayesian hierachical framework that allows to integrate environmental and anthropogenic processes into a single model.*

Keywords. *Bayesian modelling; Ecology; Spatial and spatio-temporal covariance modelling*

---

1 Introduction

Marine litter, defined as any synthetic materials lost, discarded or transported in the marine environment [1], has become a growing threat that might jeopardize the status of marine ecosystems at global and regional scale. The sources of marine debris are mainly related to human activities and include either land- and sea-based origins. In fact, in addition to pollution from ships and fishing activities, these materials enter the marine environment by rivers, drainage, sewage systems or by wind. The debris quantity and its distribution on the Mediterranean seafloor are still not well-known although preliminary studies suggest how its presence may heavily affect populations, trophic interactions and assemblages of marine living communities [2]. Despite the lack of marine litter data based on systematic monitoring/evaluating campaigns, experimental bottom trawl surveys carried out in the Mediterranean basin in the last years represent a valuable source of information about the spatial distribution and the composition of wastes. Litter typologies can be seen as special items caught by the trawl net together with marine species. While single-species distribution models have been commonly used to explain and predict the response of different taxa to environmental variation, the analysis at the community-level is still lacking [7]. Distance-based ordination methods provide insights to describe patterns of diversity and community composition, but they are deficient in explaining the relative contributions across space and time. Some innovative approaches that explicitly acknowledge the multivariate nature of species assemblages were

recently proposed [3, 7]. These approaches share the possibility to model the actual processes that determine the assemblage of community samples, taking into account for the various sources of correlation across species. In this study we analyze multivariate litter abundance data using a correlated response models in the spirit of [4]. This model merges univariate Generalized Linear Models with latent variables to account for the residual correlation across litter categories, *e.g.* due to environmental interactions or unaccounted covariates. Latent variables provide a method for "residual ordination". They are included alongside the measured covariates and are interpreted as a device to account for any residual covariation not explained by the environmental covariates. The whole implementation is performed in a hierarchical Bayesian framework, that proves to be flexible enough to integrate many data generating processes into a single model.

2 Materials and methods

2.1 Study area and data

Litter data are collected during experimental trawl surveys conducted from 2013 to 2017 in the North-Western Ionian Sea as a complementary (voluntary) activity of the international project MEDITS (MEDiterranean International Trawl Surveys). The study area (GSA 19) covers a total surface of 16,350 km^2 at depths between 10 and 800 m. The North-Western Ionian is the deepest sea in the Mediterranean basin and is characterised by a complex geomorphology and divided in two sectors by the Taranto Valley: an Eastern sector between the Taranto Valley and the Apulia represented by a broad continental shelf; a South-Western one (along the Calabria and Sicily) with a very limited shelf and many submarine canyons located along these coasts, playing an important role in the transport of terrigenous debris from coastal waters to deeper grounds. In the North-Western Ionian Sea fishing occurs from coastal waters to about 800 m with Gallipoli, Taranto, Crotone and Reggio Calabria representing the most important fisheries as well as the main harbours. Moreover, an increasing touristic activity is developing along the Ionian coasts. So, the sea bottoms are here exposed to a strong increase in anthropogenic impact both through extension of the coastal fisheries to the slope, and due to other coastal and offshore activities. The same 70 depth-stratified hauls are carried out between 10 and 800 m in depth every year (Figure 1A), summing to 350 hauls in 5 years. Wastes caught during the trawl surveys are classified in 8 categories: plastic, rubber, metal, glass/ceramic, cloth/natual fibres, processed wood, paper/cardboard, other/unspecified. The number of collected items for each litter category was scaled to the swept surface unit (1 km^2), thus obtaining density indices (N/km^2) for each litter category and survey at every haul location. Litter density is a semi-continuous zero-inflated non-negative variable. Preliminarily, to investigate factors influencing the density of litter categories, we consider the depth of the haul as environmental covariate.

2.2 Model-based statistical framework

Densities of litter categories are jointly modeled as semi-continuous zero-inflated multivariate responses assuming the Tweedie distribution model [6]. The mean density μ_{ij} of j-th litter category at the i-th haul is specified by the following mixture model:

$$g(\mu_{ij}) = \alpha_1(\boldsymbol{t}_i) + \alpha_2(\boldsymbol{s}_i) + \boldsymbol{\beta}_{0j} + \sum_{k=1}^{p} \boldsymbol{\beta}_{jk}\boldsymbol{X}_{ik} + \boldsymbol{z}_i'\boldsymbol{\theta}_j \qquad i = 1,\ldots,350; \quad j = 1,\ldots,8 \tag{1}$$

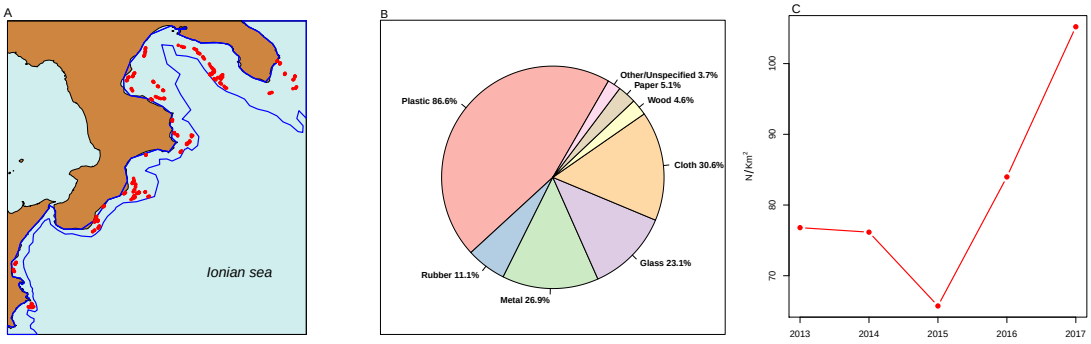


Figure 1: (A) Map of the study region and haul locations; (B) Overall percentages of litter categories; (C) Overall temporal trend of the litter density

where $g(\cdot)$ is the link function, $\alpha_{1,2}(\cdot)$ are effects adjusting for differences in site and time (year) on the overall litter density, β_{0j} is the litter type-specific intercept and β_{jk} is the type-specific regression coefficient of the k-th covariate (preliminarily, only the *depth* covariate was considered in this work). Finally, $\boldsymbol{z_i} = (z_{i1}, \ldots, z_{iq})'$ is a q-dimensional vector of latent variables, while $\theta_j = (\theta_{j1}, \ldots, \theta_{jq})$ are the corresponding litter type-specific loadings. Independent weakly informative $N \sim (0, 10)$ priors were assumed for all site and time effects, type-specific intercepts, type-specific regression coefficients, latent variables and loadings. Uniform priors $U \sim (0, 30)$ are adopted for all dispersion and variance parameters in the model. Inferences for model in Eq. 1 were implemented by the `boral` package [5] that provides an interface between `R` and `JAGS` [8] for multi-species models with latent variables.

3 Results

Figure 1B reports the overall percentages of litter categories for the study period. As expected, plastic is the prevalent litter category found in almost 90% of considered hauls. The overall density for all litter categories shows a negative peak in 2015 with a clear increasing trend for the following two years (Figure 1C). The model in Eq.1 was fitted with 1 to 3 latent variables and with fixed or random site and time effects. All model estimates were obtained using 20,000 iterations, discarding the first 5,000 corresponding to the *burn-in* phase of the algorithm. The Geweke convergence diagnostic and the graphical inspection of the trace provided clear evidence of the convergence of MCMC chains for all model parameters. As reported in Table 1, the best models in terms of lowest BIC consider random site/year effects instead of fixed effects. Here we prefer to report results from the model with two latent variables as it enables to draw a scatterplot of the ordinations, in line with distance-based techniques where two axes are typically chosen for low-dimensional data visualization (Figure 2). Figures 3A-B show a positive correlation between plastic and glass litter due to depth. Strong, positive residual correlations are observed: the plastic is correlated with all other materials except for metal and other/unspecified wastes, as also shown in Figure 2. Finally, estimated spatial effects represented in Figure 3C, allow to identify some "hot-spots" assemblages for all litter categories.

Acknowledgments. C. Calculli and A. Pollice were supported by the PRIN2015 project "Environmental processes and human activities: capturing their interactions via statistical methods (EPHASTAT)" funded by MIUR - Italian Ministry of University and Research.

	Site/year effect	
	fixed	random
1 LV	8731.59	8387.27
2 LV	8777.11	8434.66
3 LV	8833.30	8486.78

Table 1: Values of the BIC for models with 1-3 latent variables (LV) and fixed/random time and site effects

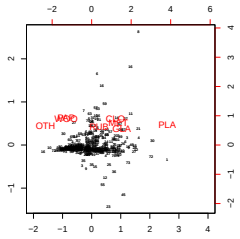


Figure 2: Residual ordination biplot for the litter densities based on the posterior median estimates

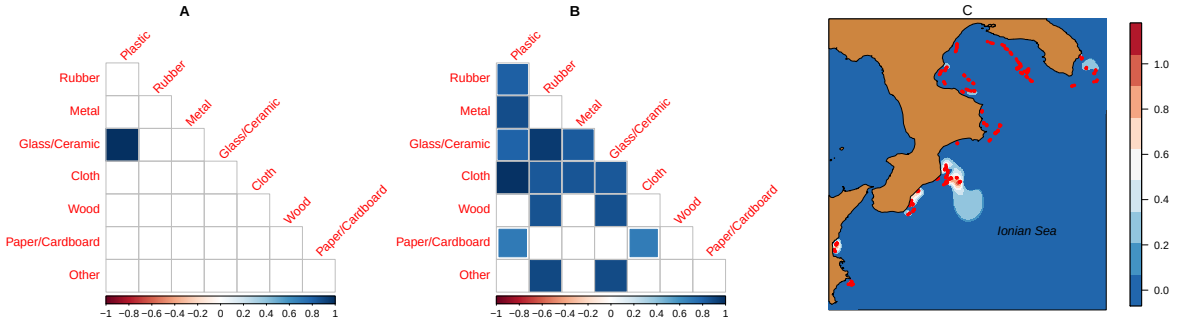


Figure 3: (A) Correlations between litter categories due to the depth environmental covariate; (B) Residual correlations based on the correlated response model (significant correlations based on 95% credible intervals excluding zero, have been reported); (C) Predictions of random spatial effects of sites.

References

[1] Galgani, F., Fleet, D., Van Franeker, J., Katsanevakis, S., Maes, T., Mouat, J., Oosterbaan, L., Poitou, I., Hanke, G., Thompson, R., Amato, E., Birkun, A., Janssen, C. (2010). Marine Strategy Framework Directive, Task Group 10 Report: Marine Litter. In JRC Scientific and Technical Reports (ed. N. Zampoukas). Ispra: European Commission Joint Research Centre.

[2] Gall, S.C., Thompson,R.C. (2015). The impact of debris on marine life. *Marine Pollution Bulletin* **92**, 170–179.

[3] Hui, F.K.C., Taskinen, S., Pledger, S., Foster, S. D., Warton, D. I. (2015). Model–based approaches to unconstrained ordination. *Methods in Ecology and Evolution* **6**, 399–411.

[4] Hui, F. K. C. (2016). boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in R. *Methods in Ecology and Evolution* **7**, 744–750.

[5] Hui, F. K. C. (2017). boral: Bayesian Ordination and Regression AnaLysis. R package version 1.4.

[6] Jørgensen, B. (1997). The Theory of Dispersion Models. Chapman and Hall. London

[7] Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T. and Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letter* **20**, 561–576.

[8] Plummer, M. et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003) March*, 20–22. Vienna, Austria.

CircSpaceTime: an R package for spatial and spatio-temporal modeling of Circular data

Mario Santoro[1,2], Gianluca Mastrantonio[3] and Giovanna Jona Lasinio[1,*]

[1] *Department of Statistical Sciences, University of Rome "Sapienza"; mario.santoro@uniroma1.it, giovanna.jonalasinio@uniroma1.it*
[2]*Istituto per le Applicazioni del Calcolo "Mauro Picone" - CNR*
[*]*Corresponding author*
[3] *Polytechnique of Turin; gianluca.mastrantonio@polito.it*

Abstract. *CircSpaceTime is going to be a new R package that eventually will implement most of the models recently developed for spatial and spatio-temporal interpolation of circular data. Such data are often found in applications where, among the many, wind directions, animal movement directions, and wave directions are involved. To analyze such data we need models for observations at locations* **s** *and times t, so-called geostatistical models providing structured dependence which is assumed to decay in distance and time. For example for wave directions in a body of water, we imagine a wave direction at every location and every time. Thus, the challenge is to introduce structured dependence into angular data. The approach we take begins with models for linear variables over space and time using Gaussian processes. Then, we use either wrapping or projection to obtain Gaussian processes for circular data. Altogether, this package will implement in its first version the proposals by Mastrantonio, Jona Lasinio and Gelfand. The models are cast as hierarchical models, with fitting and inference within a Bayesian inference framework. All procedures are written using Rcpp and whenever possible, the computation is parallelized. We use a wave direction dataset as a running example.*

Keywords. *Circular data; Bayesian modeling; Computational statistics; R packages.*

---

1 Overview

CircSpaceTime is an R package implementing spatial and spatio-temporal models for circular data, yet to be released. As a first step we start with a purely spatial setting, as initially introduced in [12].
There are different approaches to specify valid circular distributions, see for example [11]. Here we focus on two methods that allow to build a circular distribution starting from a linear one, namely the wrapping, and the projection. Under both methods, the resulting distribution has a complex functional form but introducing a suitable latent variable, the joint distribution of observed and latent variables are easy to handle in a fully Bayesian framework.

2 Defining circular processes

The wrapping approach Let $Y \in \mathbb{R}$ be random variable defined on the real line (*linear random variable*) with probability density function (pdf) $f_Y(\cdot|\psi)$, where ψ is a generic vector of parameters. We can obtain a circular random variable using the following transformation:

$$\Theta = Y \bmod 2\pi \in [0, 2\pi). \tag{1}$$

The pdf of Θ is

$$f_\Theta(\theta|\psi) = \sum_{k=-\infty}^{\infty} f_Y(\theta + 2\pi k|\psi). \tag{2}$$

Between Y and Θ there is the following relation: $Y = \Theta + 2\pi K$, where K is called the *winding number*. Equation (2) wraps $f_Y(\cdot|\psi)$ around the unit circle and Θ is called the *wrapped* version of $\mathbf{Y}$ with period 2π, e.g. if Y is normally distributed, then Θ follows a *wrapped normal* (WN) distribution. Notice that the wrapping approach preserves many properties of the distribution of the linear variable, for example continuity at the origin: $f_\Theta(0|\psi) = \sum_{k=-\infty}^{\infty} f_Y(0 + 2\pi k|\psi) = \sum_{k=-\infty}^{\infty} f_Y(2\pi + 2\pi k|\psi)$, we have then that $\sum_{k=-\infty}^{\infty} f_Y(2\pi + 2\pi k|\psi) = \lim_{\theta \to 2\pi} \sum_{k=-\infty}^{\infty} f_Y(\theta + 2\pi k|\psi)$ as long as $f_Y(\theta + 2\pi k|\psi)$ is continuous in $2\pi k$ for all k. It is not easy to work directly with equation (2), since it requires the evaluation of an infinite sum. Following [6], if we consider K as (latent) random variable we can see that $f_{\Theta,K}(\theta, k|\psi) = f_Y(\theta + 2\pi k|\psi)$, i.e. $f_Y(\theta + 2\pi k|\psi)$ is the joint density of (Θ, K), and a marginalization over K gives equation (2). The conditional distribution of K, needed for the implementation of the MCMC, is easy to handle since it is proportional to : $f_Y(\theta + 2\pi k|\psi)$. It is then generally easier to work with the joint density of $\Theta, K|\psi$, with respect to the one of $\Theta|\psi$, since the former does not require the evaluation of the infinite sum.
The wrapping approach can be easily extended to a multivariate setting [12]. Let $\mathbf{Y} = (Y_1, \ldots, Y_p)'$ be a p-variate vector with pdf $f_\mathbf{Y}(\cdot|\psi)$, then $\mathbf{\Theta} = (\Theta_1, \ldots, \Theta_p)'$, with $\Theta_i = Y_i \bmod 2\pi \in [0, 2\pi)$ that is a vector of circular variables. Extending the univariate approach, we can easily find that the full conditional of $\mathbf{K}$ is proportional to $f_\mathbf{Y}(\boldsymbol{\theta} + 2\pi\mathbf{k}|\psi)$ and the joint density of $(\mathbf{\Theta}, \mathbf{K})$ has a multivariate functional form. Here again it is easier to treat $\mathbf{K}$ as a latent variable.

The projection approach Let $\mathbf{Y} = (Y_1, Y_2)$ be a bivariate vector of linear variables with pdf $f_\mathbf{Y}(\cdot|\psi)$. The unit vector $\mathbf{U} = \frac{\mathbf{Y}}{||\mathbf{Y}||}$ represents a point over the unit circle and the associated angle Θ, where $U_1 = \cos(\Theta)$ and $U_2 = \sin(\Theta)$, is a circular random variable; we have then $\tan(\Theta) = \frac{Y_2}{Y_1} = \frac{U_2}{U_1}$. Since the period of the tangent is π, inversion of this function, to obtain Θ, requires some care. A common choice is the atan*, formally defined in [11], pag. 13, that takes into account the signs of Y_1 and Y_2 to determine the right portion of the unit circle where Θ is located. Between Θ and $\mathbf{Y}$ the following relation exists $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = R \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} = R\mathbf{U}$, with $R = ||\mathbf{Y}||$. The pdf of $\Theta|\psi$ is $f_\Theta(\theta|\psi) = \int_{\mathbb{R}^+} r f_\mathbf{Y}((r\cos(\theta), r\sin(\theta))'|\psi) dr$. The integral in this equation is not easy to solve and, even when a closed form exists, the resulting pdf has a complicated functional structure. The joint density of $\Theta, R|\psi$ is $f_\mathbf{Y}((r\cos(\theta), r\sin(\theta))|\psi)$, and if $\mathbf{Y} \sim N_2(\boldsymbol{\mu}, \mathbf{\Sigma})$ then $f_{\Theta,R}(\theta, r|\psi) = r\phi_2((r\cos(\theta), r\sin(\theta))'|\boldsymbol{\mu}, \mathbf{\Sigma})$.
The projection approach can be easily adapted to obtain a distribution for multivariate circular variable [16]. If $\mathbf{Y}$ is a $2p-$variate linear variable, a $p-$variate vector of (projected) circular variables is obtained with the following transformation: $\Theta_i = \text{atan}^*\left(\frac{Y_{2i}}{Y_{2i-1}}\right)$, $i = 1, \ldots, p$. The pdf of $\mathbf{\Theta}|\psi$, where $\mathbf{\Theta} = (\Theta_1, \ldots, \Theta_p)'$, is

$$f_\mathbf{\Theta}(\boldsymbol{\theta}|\psi) = \int_{\mathbb{R}^+} \cdots \int_{\mathbb{R}^+} \prod_{i=1}^{p} r_i f_\mathbf{Y}(\mathbf{y}|\psi) dr_1 \ldots dr_p, \tag{3}$$

where $r_i = ||(y_{2i-1}, y_{2i})'||$ and, in equation (3), $\mathbf{y}$ is a function of $\boldsymbol{\theta}$ and $\mathbf{r} = (r_1, \ldots, r_p)$.
In the multivariate case, as in the univariate one, it is generally easier to work with the joint density of $\boldsymbol{\Theta}, \mathbf{R}|\boldsymbol{\psi}$ (the integrand in (3)) than the one of $\boldsymbol{\Theta}|\boldsymbol{\psi}$.

Spatio-temporal processes for circular variables A stochastic process can be defined through its finite dimensional distribution, i.e. the distribution of an $n-$dimensional realization, that is a multivariate pdf [10]. Starting from a distribution for linear variables, we can use the wrapping or the projection approach to obtain circular distributions. Then from an $n-$dimensional realization of a linear process, we can obtain an $n-$dimensional realization of a circular one. More precisely, let $\mathbf{Y}(\mathbf{s}) \in \mathbb{R}^p$, with $\mathbf{s} \in \mathcal{S} \subset \mathbb{R}^d$, be a $p-$variate stochastic process, defined over a $d-$dimensional domain, and suppose that an $n-$dimensional realization of the process $\mathbf{Y}(\mathbf{s})$, $\mathbf{y}$, has pdf $f_{\mathbf{Y}}(\cdot|\boldsymbol{\psi})$.
Wrapped circular process Let $p = 1$ and let $\mathbf{y} = (y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n))$ be the $n-$dimensional realization of $\mathbf{Y}(\mathbf{s})$. If we apply the transformation (1) to each component of $\mathbf{y}$, we obtain a vector of dimension n of wrapped circular variables: $\boldsymbol{\theta} = (\theta(\mathbf{s}_1), \ldots, \theta(\mathbf{s}_n))'$. [12] show that the vector $\boldsymbol{\theta}$ is the $n-$dimensional realization of the circular process $\Theta(\mathbf{s}) = \mathbf{Y}(\mathbf{s}) \bmod 2\pi$, with $\mathbf{Y}(\mathbf{s}) = \Theta(\mathbf{s}) + 2\pi\mathbf{K}(\mathbf{s})$.
Projected circular process Let $p = 2$, then $\mathbf{Y}(\mathbf{s}) = (Y_1(\mathbf{s}), Y_2(\mathbf{s}))'$ is a bivariate process and $\mathbf{y} = (\mathbf{y}(\mathbf{s}_1), \ldots, \mathbf{y}(\mathbf{s}_n))'$, its finite-dimensional realization, is a vector of bivariate variables, i.e. $\mathbf{y}(\mathbf{s}_i) = (y_1(\mathbf{s}_i), y_2(\mathbf{s}_i))' \in \mathbb{R}^2$. The projected circular process is obtained as $\Theta(\mathbf{s}) = \text{atan}^*\left(\frac{Y_2(\mathbf{s})}{Y_1(\mathbf{s})}\right)$, i.e. we apply this transformation to the process $\mathbf{Y}(\mathbf{s})$ [16]. The finite dimensional realization of the circular process is $\boldsymbol{\theta} = (\theta(\mathbf{s}_1), \ldots, \theta(\mathbf{s}_n))'$, where $\theta(\mathbf{s}_i) = \text{atan}^*\left(\frac{y_2(\mathbf{s}_i)}{y_1(\mathbf{s}_i)}\right), i = 1, \ldots, n$.
For both processes, we assume $\mathbf{Y}(\mathbf{s}) = \boldsymbol{\mu} + \boldsymbol{\omega}(\mathbf{s}) + \boldsymbol{\varepsilon}(\mathbf{s})$. where with $\boldsymbol{\mu}$ is a mean term, $\boldsymbol{\omega}(\mathbf{s})$ is a Gaussian process and $\boldsymbol{\varepsilon}(\mathbf{s})$ is the nugget effect. For the wrapped approach $\mathbf{Y}(\mathbf{s})$ is a univariate process, while under the projected it is bivariate.

3 Implementation details

Model parameters are estimated using a MCMC algorithm involving Gibbs sampler and, when necessary, we adopt a Metropolis within Gibbs step. It is well known that the MCMC tends to mix really slow ([1]) when latent variable are involved. To speed up the convergence, we try to find an optimal proposal distribution for the Metropolis step using the algorithm described in [15], page 258. With the goal of speeding up the MCMC convergence, is suggested to decrease the dimension of the parameters space, that is, do as much marginalization as possible ([2]).The core estimation is based, mostly, on loops with thounsands iterations. To improve performances [17] we implemented everything in *C++* and using `Rcpp` package we can simplify the integration between C++ and R codes [8]. In particular we used the `RcppArmadillo` package [9] that implement the Armadillo matrix library for it simplicity and elegance [7], although the `RcppEigen` is a bit faster [3]. For a fast multiple chain estimations we used `doParallel` package [14].

4 Conclusions and future developments

At present the package implements the spatial wrapped Gaussian and the Projected Gaussian processes. The authors hope to be able to present the spatio-temporal processes at the conference.
The next step will be to include the Hidden Markov models approach proposed in [13, 4, 5] in a set of

papers published over the last few years focusing on spatial and spatio-temporal modeling and classifications. These set of models will be estimated in a likelihood approach as originally proposed by their authors.

Acknowledgements The authors are partially supported by the MIUR-PRIN grant EphaStat (20154X8K23-SH3).

References

[1] Andrieu, C., Doucet, A. and Holenstein, R. (2010) , 'Particle Markov chain Monte Carlo methods', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(3), 269–342.

[2] Banerjee, S., Gelfand, A. E. and Carlin, B. P. (2014) , *Hierarchical Modeling and Analysis for Spatial Data*, second edn, Chapman and Hall/CRC, New York.

[3] Bates, D. and Eddelbuettel, D. (2013) , 'Fast and elegant numerical linear algebra using the rcppeigen package', *Journal of Statistical Software, Articles* **52**(5), 1–24.

[4] Bulla, J., Lagona, F., Maruotti, A. and Picone, M. (2012) , 'A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series', *Journal of Agricultural, Biological, and Environmental Statistics* **17**(4), 544–567.

[5] Bulla, J., Lagona, F., Maruotti, A. and Picone, M. (2015) , 'Environmental conditions in semi-enclosed basins: A dynamic latent class approach for mixed-type multivariate variables', *Journal de la Société Française de Statistique* **156**(1), 114–137.

[6] Coles, S. and Casson, E. (1998) , 'Extreme value modelling of hurricane wind speeds', *Structural Safety* **20**(3), 283 – 296.

[7] Eddelbuettel, D. (n.d.) , 'Higher-performance r programming with c++ extensions'. Zurich R Courses 2017.

[8] Eddelbuettel, D. and Francois, R. (2011) , 'Rcpp: Seamless R and C++ integration', *Journal of Statistical Software, Articles* **40**(8), 1–18.

[9] Eddelbuettel, D. and Sanderson, C. (2014) , 'Rcpparmadillo: Accelerating R with high-performance C++ linear algebra', *Computational Statistics and Data Analysis* **71**, 1054 – 1063.

[10] Gelfand, A., Diggle, P., Fuentes, M. and Guttorp, P. (2010) , *Handbook of Spatial Statistics*, Chapman and Hall.

[11] Jammalamadaka, S. R. and SenGupta, A. (2001) , *Topics in Circular Statistics*, World Scientific, Singapore.

[12] Jona Lasinio, G., Gelfand, A. E. and Jona Lasinio, M. (2012) , 'Spatial analysis of wave direction data using wrapped Gaussian processes', *Annals of Applied Statistics* **6**(4), 1478–1498.

[13] Lagona, F., Picone, M., Maruotti, A. and Cosoli, S. (2015) , 'A hidden Markov approach to the analysis of space-time environmental data with linear and circular components', *Stochastic Environmental Research and Risk Assessment* **29**(2), 397–409.

[14] MicrosoftCorporation and Weston, S. (2017) , *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.11.

[15] Robert, C. P. and Casella, G. (2009) , *Introducing Monte Carlo Methods with R (Use R)*, 1st edn, Springer-Verlag, Berlin, Heidelberg.

[16] Wang, F. and Gelfand, A. E. (2014) , 'Modeling space and space-time directional data using projected Gaussian processes', *Journal of the American Statistical Association* **109**(508), 1565–1580.

[17] Wickham, H. (2015) , *Advanced R*, Chapman & Hall/CRC The R Series, CRC Press.

Nonparametric approach for spatial prediction incorporating information from correlated auxiliary variables

P. García-Soidán[1,*] and T. R. Cotos-Yáñez[1]

[1]*Dept. of Statistics and OR, University of Vigo (Spain); pgarcia@uvigo.es, cotos@uvigo.es*
[*]*Corresponding author*

Abstract. *Prediction of a spatial variable at an unsampled location can be addressed through the kriging methodology. Furthermore, the resulting predictor can incorporate data from secondary variables, correlated with the target one, when using cokriging techniques. These procedures demand characterizing the multivariate dependence structure, which is not an easy task. To simplify the previous issue, a nonparametric kernel predictor will be introduced in the current work, designed to include data from the target variable and auxiliary ones. The asymptotic unbiasedness of the new approach will be proved, together with the negligibility of the mean squared prediction error for large samples. A bootstrap mechanism will be proposed for approximation of the aforementioned error, adapted to resample data from a multivariate process. For the choice of the bandwidth parameters involved, balloon selectors will be suggested. We will also deal with the estimation of the remaining unknown terms in the kernel predictor, by proceeding through parametric and nonparametric techniques. Simulation studies will be developed in different scenarios to check the performance of the new methodology for prediction. Finally, the nonparametric approach will be applied to a real data set to illustrate its practical implementation.*

Keywords. *Bandwidth parameter; Cokriging; Covariogram; Kernel method; Prediction.*

1 Introduction

The kriging methodology allows the prediction of a spatial variable at an unsampled location from the available data [1, 5]. In particular, when auxiliary variables have also been observed, correlated with the target one, the cokriging techniques [11] may be applied, with the advantage that they have been designed to incorporate the information provided by the whole data set, aiming at predicting the value of the main variable. A brief summary of this methodology will be provided below.

Let us assume that $\{Z(\mathrm{s}) = (Z_1(\mathrm{s}), ..., Z_p(\mathrm{s}))/\mathrm{s} \in D \subset \boldsymbol{R}^d\}$ is a p-variate random process, so that Z_i can be modeled as $Z_i(\mathrm{s}) = \mu_i(\mathrm{s}) + Y_i(\mathrm{s})$, for $i = 1, ..., p$, where μ_i is the deterministic trend and Y_i is a second-order stationary random process with zero mean. Then, it follows that:

(i) $E[Z_i(\mathrm{s})] = \mu_i(\mathrm{s})$, for all $\mathrm{s} \in D$.

(ii) $Cov[Z_i(\mathrm{s}), Z_{i'}(\mathrm{s}')] = C_{i,i'}(\mathrm{s} - \mathrm{s}')$, for all $\mathrm{s}, \mathrm{s}' \in D$ and for all i, i'.

For $i \neq i'$, $C_{i,i'}$ represents the cross-covariogram between Z_i and $Z_{i'}$, whereas $C_{i,i}$ is the direct covariogram of Z_i. From **(ii)**, one has that $Var[Z_i(\mathrm{s})] = C_{i,i}(0) = \sigma_i^2$, for all s and i.

Z_1 will be taken as the variable of interest, while the remaining ones will be considered as secondary variables. Then, suppose that Z_i is observed on the set $\mathrm{S}_i = \{\mathrm{s}_{i,1}, ..., \mathrm{s}_{i,n_i}\}$ of $n_i > 0$ locations, for $i = 1, ..., p$. A cokriging predictor for Z_1 at an unsampled location s can be constructed as follows:

$$\hat{Z}_1(\mathrm{s}) = \sum_{i,j_i} \lambda_{i,j_i} Z_i(\mathrm{s}_{j_i}) \tag{1}$$

where the parameters λ_{i,j_i} would be derived by minimizing $E\left[\left(\hat{Z}_1(\mathrm{s}) - Z_1(\mathrm{s})\right)^2\right]$, subject to the unbiasedness condition, namely, $\mu_1(\mathrm{s}) = E\left[\hat{Z}_1(\mathrm{s})\right] = \sum_{i,j_i} \lambda_{i,j_i} E\left[Z_i(\mathrm{s}_{j_i})\right]$.

The resulting λ_{i,j_i} are dependent on functions $C_{i,i'}$, whose estimation is not an easy task [6]. Indeed, $\frac{p^2+p}{2}$ covariograms must be characterized, satisfying some constraints, and the adequateness of the fitted functions may not be properly checked through the cross-validation techniques [9]. In addition, the data correlation level, among other factors, affects the accuracy of the predicted values [12].

In view of the above, in the current research, we propose a nonparametric alternative for prediction in this setting, designed to avoid the aforementioned problems. Our idea generalizes the nonparametric approach [8], based on predicting the value of the target variable at s from the data collected for Z_1 as given below:

$$\hat{Z}_1(\mathrm{s}) = \sum_{j_1} \frac{K_d\left(\frac{\mathrm{s}-\mathrm{s}_{1,j_1}}{h(\mathrm{s})}\right)}{\sum_{j_1} K_d\left(\frac{\mathrm{s}-\mathrm{s}_{1,j_1}}{h(\mathrm{s})}\right)} Z_1(\mathrm{s}_{1,j_1})$$

where K_d represents a d-variate symmetric kernel function and $h(\mathrm{s}) > 0$ is the bandwidth parameter, for each s.

Now, we suggest incorporating the whole observed values through a weighted average, whose weights account for the correlations between the target variable and each of the secondary ones in the following manner:

$$\hat{Z}_1(\mathrm{s}) = \sum_{i,j_i} p_{i,j_i}(\mathrm{s}) \left(\mu_1(\mathrm{s}_{i,j_i}) + sign\left(C_{1,i}(0)\right) \frac{\sigma_1}{\sigma_i} \left(Z_i(\mathrm{s}_{i,j_i}) - \mu_i(\mathrm{s}_{i,j_i})\right)\right) \tag{2}$$

with:

$$p_{i,j_i}(\mathrm{s}) = \frac{K_1\left(\frac{C_{1,i}(0)^2 - \sigma_1^2\sigma_i^2}{h_i}\right) K_d\left(\frac{\mathrm{s}-\mathrm{s}_{i,j_i}}{h(\mathrm{s})}\right)}{\sum_{i,j_i} K_1\left(\frac{C_{1,i}(0)^2 - \sigma_1^2\sigma_i^2}{h_i}\right) K_d\left(\frac{\mathrm{s}-\mathrm{s}_{i,j_i}}{h(\mathrm{s})}\right)}$$

where $sign(x)$ stands for the sign of a real value x and h_i denotes a bandwidth parameter, for each i. To avoid the border effects in the above predictor, boundary kernels could be used instead of symmetric ones.

Predictor (2) satisfies good properties, as it is asymptotically unbiased and the mean-squared prediction error converges to zero as the sample sizes n_i increase, under certain hypotheses. For instance, if a random design is assumed for the spatial locations, the underlying density should be sufficiently smooth, although the validity of the aforementioned properties could also be extended to deterministic designs. In addition, the covariograms $C_{i,i'}$ will be required to admit a number of derivatives in a neighborhood of 0. However, specification of the dominant terms of the mean-squared prediction error becomes rather cumbersome. Thus, we propose using instead a bootstrap approach for approximation of this error in practice, as the one introduced in [2], but adapted to resample from multivariate data.

For selection of the bandwidth $h(\mathrm{s})$, assume that the kernel density K_d is compactly supported, namely, that $K_d(\mathrm{y}) = 0$, for all $\mathrm{y} \in \boldsymbol{R}^d$ such that $\|\mathrm{y}\| \geq a_d$, for some $a_d > 0$. Then, $h(\mathrm{s})$ could be selected by using the balloon approach [10] and it can be taken as the m-th percentile (for instance, m equal to 10% or 20%) of the values $a_d^{-1}\|\mathrm{s} - \mathrm{s}_{i,j_i}\|$, for each s.

Regarding the choice of h_i, suppose that the correlation is considered significant if $Corr[Z_1(\mathrm{s}), Z_i(\mathrm{s})]^2 > b$, for some b such that $0 < b < 1$ (for instance, b equaling 0.8 or 0.9). The balloon estimation would enable the selection of h_i as $a_1^{-1}\sigma_1^2\sigma_i^2(1-b)$.

Implementation of predictor (2) also requires estimation of the remaining unknown terms, dependent on the trend and on the covariogram functions. These issues can be addressed in a variety of ways, according to the hypotheses that are assumed from the underlying random process. In addition, parametric or nonparametric approaches can be used to derive the necessary estimates. Application of the parametric methodology is easier but it can be affected by misspecification of the selected models. Thus, nonparametric alternatives will also be suggested instead, which can be even employed in practice in combination with the parametric approaches.

The new prediction methodology must provide accurate results for gaussian data, similarly as could be expected for the cokriging predictor (1). However, the latter one demands characterization of the whole direct and cross-covariograms. In addition, the resulting functions must be appropriately fit to be valid for prediction, as well as satisfy a number of restrictions due to the underlying dependence among them, unlike what happens to predictor (2). Indeed, our proposal only involves terms dependent on the covariogram functions of the form $C_{1,i}(0)$ or $C_{i,i}(0) = \sigma_i^2$, thus simplifying their practical implementation. This aim can be addressed through a parametric fit [4] or a nonparametric mechanism, based on the method of moments [3] or on the application of the kernel method [7].

Under isotopy, predictor (2) can be expressed in a simpler way. With this aim, consider that Z_i is observed on the set $\mathrm{S}_i = \mathrm{S} = \{\mathrm{s}_j\}_{j=1}^n$ of n locations, for all $i = 1, ..., p$. The kernel-type predictor of Z_1 at location s would then be given by:

$$\hat{Z}_1(\mathrm{s}) = \sum_j p_{1,j}(\mathrm{s})\tilde{Z}_1(\mathrm{s}_j)$$

with:

$$\tilde{Z}_1(\mathrm{s}_j) = \mu_1(\mathrm{s}_j) + \sum_i p_{2,i}\left(sign\left(C_{1,i}(0)\right)\frac{\sigma_1}{\sigma_i}\left(Z_i(\mathrm{s}_j) - \mu_i(\mathrm{s}_j)\right)\right)$$

$$p_{1,j}(\mathrm{s}) = \frac{K_d\left(\frac{\mathrm{s}-\mathrm{s}_j}{h(\mathrm{s})}\right)}{\sum_j K_d\left(\frac{\mathrm{s}-\mathrm{s}_j}{h(\mathrm{s})}\right)},\ p_{2,i} = \frac{K_1\left(\frac{C_{1,i}(0)^2 - \sigma_1^2\sigma_i^2}{h_i}\right)}{\sum_i K_1\left(\frac{C_{1,i}(0)^2 - \sigma_1^2\sigma_i^2}{h_i}\right)}$$

Acknowledgments. The first author acknowledges financial support from the Spanish National Research and Development Program project [TEC2015-65353-R], from the ERDF and from the Xunta de Galicia (Spain) under project GRC2015-018 and under agreement for funding AtlantTIC (Atlantic Research Center for Information and Communication Technologies). The second author's work has been partially supported by the Grant MTM2017-89422-P of the Spanish Ministry of Science and Innovation and by the Xunta de Galicia (Spain) under project ED431C2016-040(GRC).

References

[1] Cressie, N. (1993). *Kriging for spatial data*. Wiley. New York.

[2] García-Soidán, P., Menezes, R., Rubiños López, O. (2014). Bootstrap Approaches for Spatial Data. *Stochastic Environmental Research and Risk Assessment* **28**, 1207–1219.

[3] Genton, M. G., Kleiber, W. (2015). Cross-covariance functions for multivariate geostatistics. *Statistical Science* **30**, 147–163.

[4] Gneiting, T., Kleiber, W., Schlather, M. (2012). Matérn Cross-Covariance Functions for Multivariate Random Fields. *Journal of the American Statistical Association* **105**, 1167–1177.

[5] Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press. Oxford.

[6] Goulard, M., Voltz, M. (1992). Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix. *Mathematical Geology* **24**, 269–282.

[7] Kleiber, W., Nychka, D. (2012). Nonstationary modeling for multivariate spatial processes. *Journal of Multivariate Analysis* **112**, 76–91.

[8] Menezes, R., García-Soidán, P., Ferreira, C. (2010). Nonparametric spatial prediction under stochastic sampling design. *Journal of Nonparametric Statistics* **22**, 363–377.

[9] Solow, A. R. (1990). Geostatistical cross-validation: A cautionary note. *Mathematical Geology* **22**, 637–639.

[10] Terrell, G., Scott, D. W. (1992). Variable kernel density estimation. *Annals of Statistics* **20**, 1236–1265.

[11] Wackernagel, H. (2003). *Multivariate geostatistics: an introduction with applications*. Springer. Berlin.

[12] Yalcin, E. (2005). Cokriging and its effect on the estimation precision. *The Journal of the South African Institute of Mining and Metallurgy* **105**, 223–228.

Probability maps for extreme wildfire

M. Antónia Amaral Turkman[1,*], Kamil Feridun Turkman[2] and Paula Pereira[3,4]

[1] *Centro de Estatística e Aplicações, Universidade de Lisboa; maturkman@fc.ul.pt*
[2] *Centro de Estatística e Aplicações, Universidade de Lisboa; kfturkman@fc.ul.pt*
[3] *Escola Superior de Tecnologia de Setúbal, Instituto Politécnico de Setúbal; paula.pereira@estsetubal.ips.pt*
[4] *Centro de Estatística e Aplicações, Universidade de Lisboa;*
[*] *Corresponding author*

Abstract. *Wildfires, particularly in Portugal, are a relevant public policy issue due to the significant economical and social damage they cause. Most wildfires are extinguished upon ignition and do not have significant effect. However, it is generally an established fact that a small number of fires cause most of the damage, often expressed by the phrase "*1*% of the fires do* 99*% percent of the damage" [4]. Therefore, It is very important to understand the causes of extreme fires, their spatial distribution as well as predicting the onset of a possible extreme wildfires. Probability maps indicating where ignition is most likely to take place and the consequent fire scar sizes are very important administrative tools in managing wildfires. In this talk, we will define the concept of probability maps for wildfires, describe different types of data that are commonly available and the consequent different spatio-temporal modeling strategies using Bayesian hierarchical methods [1].*

Keywords. *Wildfires, Point processes, Spatio-temporal models, Bayesian hierarchical models.*

---

1 Introduction

Vegetation fires are inherently random in the timing of their location and occurrence, in the detailed behavior of each individual event, and in the particularities of their effects on soils, water, flora, fauna, and air. Therefore, substantial efforts have been directed towards statistical modeling of several fire-related processes ([1]). One important fire-related process is fire likelihood or fire danger, which deals with pre-fire events and aims at predicting the probability of fire occurrence and the extent of area burned over a specific spatial area and temporal period, conditional on a fire occurrence. Data sources coming from satellite images and ground sources are point referenced and therefore are more suited in understanding the spatial point patterns of fire incidences as well as fire sizes. Ideally, the data on point patterns, should be treated as a realization of a spatio-temporal marked point process, discrete in time and continuous in space. Typically such point processes are modeled by marked Poisson point processes and the non-homogeneous intensity function together with fire size distribution of this process become the focal point of the study ([2]). Typically Log Gaussian Cox processes are used for modeling point patterns whereas a variety of models, among which the generalized Pareto distribution are often employed for modeling large fire sizes and extreme value theory is the natural inferential tool to quantify the large fire danger. Further simplifications, at the cost of additional loss of information, can be achieved by transforming the point pattern data into fire incidence data. The marks, namely sizes of the individual fires, can be aggregated into burned area fraction of each areal unit during the temporal units. The spatial support for the

analysis is a regular grid, and not individual fire events and consequently large fires that spanned more than one grid cell will be subdivided and have their area distributed by the corresponding cells. Grid cell records for each year represent the binary data indicating the presence of at least one fire together with the corresponding burned area fraction of the grid cell. We call these binary data the fire incidence data. The burned area fraction, expressed in terms of the percentage of grid size burned each year, is used as covariate information. Information from fire severity and maps of state of vegetation (green/dry) and cumulated biomass at the end of spring can be used to produce annual fire risk maps as in [3], by incorporating the strong spatial and temporal dependence that exists in the data. We will consider for the purpose a Markovian structure for the fire incidence data. The objective of this model is to capture, as much as possible, the strong spatio-temporal dependence structures in the fire incidence data, allowing at the same time for the introduction of any type of dynamic explanatory variables in the model. This will be achieved through Bayesian hierarchical modeling techniques and simulation-based inference.

Fire danger can then be defined, depending on the format of the data and the consequent model used for describing the data:

(i) If one bases inferences on the fire incidence or fire frequency data over areal units, then fire danger can be defined as the probability of fire in any year and in any areal unit given the observed fire incidences up to that year.

(ii) If fire severity is included in the definition, than fire danger can be defined as a fire that ignites at a location s will have consequent fire size above a high threshold.

2 Objectives, data structure and methods

The objective of this research is to obtain fire risk maps for 2018 based on satellite data of fire incidence and fire sizes in Portugal from 1988 to 2017. For fire incidence, probability maps are constructed based on data at a grid level cell of $4km^2$. For extreme fires, probability maps are constructed at county level, based on marked point reference data.

All the models were run using R-INLA ($http://www.r-inla.org/$)

2.1 Markov model for fire incidence data

Let T be the number of years under study and N the total number of grid cells. Let $Y(i,t), t = 1,...,T, i = 1,...,N$, represent the indicator variable of fire incidence. We assume a Markovian structure for Y as in [3] where the spatial dependence is introduced in the link functions for the transition probabilities through an ICAR model. For each cell, percent of forest and shrubland cover and a cumulative Daily Severity Rating (DSR, a meteorological rating for assessing the risk of fires by using the forest fire index) obtained at the end of June, enter as covariates. Also, for each cell, the time since last fire enters as a covariate in the linear predictor for the transition probability from state 0 (no fire) at time t to state 1 (fire) at time $t+1$, with a regionally dependent coefficient. For the linear predictor for the transition probability from state 1 (fire) at time t to state 1 (fire) at time $t+1$, we use, as covariate, the percent burned area at time t again with a regionally dependent coefficient.

2.2 Model for extreme fires

Let v be a fire size assumed extreme (say $v = 1000$ ha) and u a fire size such that excess fires above it can be modelled by a generalized Pareto distribution(GPD). The objective here is to obtain the probability that at least on fire size above v hectares will be observed in year t at a county A.

Consider the data of all the locations and fire scares above u for every year t for every county A. Let $\mathbf{s}_t$ be the set of locations of fire scares in year t and $\mathbf{x}(s_t)$ the set of excess fire sizes above u and $\mathbf{y}(s_t)$ a vector of covariates (similar covariates as used in the previous model). We assume that for every t, $(\mathbf{s}_t, \mathbf{x}(s_t))$ is a marked point process, where the points follow a log Gaussian Cox process with intensity given by $\lambda(s_t, x(s_t)) = \lambda(s_t) f_{s_t}(x)$ and the marks follow a generalized Pareto distribution with density $f_{s_t}(x)$. Hence the probability that at least one fire of size above v will be observed in year t in a region A is given by

$$1 - \exp\left\{ - \int_{s \in A} \int_{x > v} \lambda(s,x) ds dx \right\}.$$

Acknowledgments. This work is partially sponsored by national funds through FCT - Fundação para a Ciência e a Tecnologia under the project UID/MAT/00006/2013.

References

[1] Pereira J.M.C., Turkman K. F. (2018). *Statistical models of wildfires; Spatial and temporal patterns.* Handbook of Environmental and Ecological Statistics, (eds Alan E. Gelfand, Montserrat Fuentes and Jennifer A. Hoeting). Chapman and Hall, Boca Raton. (to appear).

[2] Pereira, P., K. F. Turkman, M.A. Amaral Turkman, A. Sa and J. M.C. Pereira (2013) Quantification of annual wildfire risk; A spatio-temporal point process approach. Statistica, 73, 55-68.

[3] Turkman, K.F., Amaral Turkman, M.A., Pereira,P. Sa, A., and Pereira, J.(2014) Generating annual fire risk maps using Bayesian hierarchical models. Journal of Statistical Theory and Practice. 8:3, 509-533.

[4] Strauss D., Bednar L., Mees R. (1989). Do one percent of the forest fires cause ninety-nine percent of the damage? *Forest science*, **35**, 319–328.

Nonparametric bootstrap approach for risk mapping under heteroscedasticity

S. Castillo-Páez[1], R. Fernández-Casal[2] and P. García-Soidán[3,*]

[1]*Universidad de las Fuerzas Armadas ESPE (Ecuador); sacastillo@espe.edu.ec*
[2]*Universidad de A Coruña (Spain); ruben.fcasal@udc.es*
[3]*Universidad de Vigo (Spain); pgarcia@uvigo.es*
[*]*Corresponding author*

Abstract. *The aim of this work is to provide a nonparametric resampling method for approximating the (unconditional) probability that a spatial variable exceeds a prefixed threshold value, from the available data. Then, a risk map of the target variable can be obtained, which is of great applicability in the environmental setting, for instance, to assess the contamination risk by any pollutant. Other approaches suggested for the same issue require stationarity from the random process or relax this hypothesis to admit the presence of a deterministic trend, although all of them assume constant variance throughout the observation region. However, our proposal has been designed to be valid under heteroscedasticity of the spatial process. For this purpose, local linear estimates of the trend, variance and variogram functions must be derived, where the two latter ones are corrected to reduce the bias due to the use of residuals. These estimates are employed in the implementation of a bootstrap procedure, whose replicates allow the approximation of the aforementioned risk. The performance of this mechanism is checked through numerical studies with simulated and real data.*

Keywords. *Heteroscedasticity; Local linear regression; Resampling method.*

---

1 Introduction

The current work is focused on the construction of a risk map of threshold-exceeding probabilities, from the available data. This kind of results has important applications, especially in the environmental field, where it enables the assessment of the contamination risk by any pollutant and the subsequent decision making. Thus, write $\{Y(\mathbf{x}) : \mathbf{x} \in D \subset \mathbb{R}^d\}$ for a spatial process and suppose that n data, $\mathbf{Y} = (Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n))^t$, have been collected at the respective locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Usually, the aim is the estimation of:

$$P(Y(\mathbf{x}_0) > c | \mathbf{Y}),$$

which represents the conditional probability or risk that $Y(\mathbf{x}_0)$ exceeds a fixed threshold value c, at each $\mathbf{x}_0 \in D$, for some $c > 0$. Different approaches have been provided in the statistics literature to deal with the conditional risk estimation, such as indicator kriging [6], Markov chain modeling [8] or more recent techniques, as those based on compositional data analysis [10]. Since these methods require the selection of parametric models that can be affected by misspecification, alternative options are provided by the nonparametric procedures, as the kernel-based approach proposed in [5]. However, assessment of the long-term risk [7] and other problems demand knowledge of the process distribution, under certain

general conditions. Then, the unconditional probability is needed instead:

$$r_c(\mathbf{x}_0) = P\left(Y(\mathbf{x}_0) > c\right), \tag{1}$$

and the previous methods would not be appropriate. A bootstrap geostatistical technique for the estimation of the unconditional risk was suggested in [4]. This approach can be applied to spatial processes with a deterministic trend, although it requires homoscedasticity, as all the aforementioned procedures. Thus, our research goes a step further, aiming to provide a nonparametric mechanism for approximation of the unconditional risk, which is valid for heteroscedastic spatial processes.

Suppose that the spatial process Y can be modeled as:

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + \sigma(\mathbf{x})\varepsilon(\mathbf{x}), \tag{2}$$

where μ and σ^2 denote the deterministic trend and variance functions, respectively, and ε is a second-order stationary process, with zero mean, unit variance and correlogram $\rho(\mathbf{u}) = Cov(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{x}+\mathbf{u}))$. The semivariogram of ε is given by $\gamma(\mathbf{u}) = \frac{1}{2}Var(\varepsilon(\mathbf{x}) - \varepsilon(\mathbf{x}+\mathbf{u})) = 1 - \rho(\mathbf{u})$.

The specification of the small-scale variability of the process Y requires the estimation of the variance and the correlogram (or the variogram) of the error process ε, since:

$$Cov(Y(\mathbf{x}), Y(\mathbf{x}+\mathbf{u})) = \sigma(\mathbf{x})\sigma(\mathbf{x}+\mathbf{u})\rho(\mathbf{u}).$$

Consequently:

$$\boldsymbol{\Sigma} = \mathbf{DRD},$$

where $\boldsymbol{\Sigma}$ and $\mathbf{R}$ denote the covariance matrices of $\mathbf{Y}$ and $\varepsilon = (\varepsilon(\mathbf{x}_1), \ldots, \varepsilon(\mathbf{x}_n))^t$, respectively, and $\mathbf{D} = diag(\sigma(\mathbf{x}_1), \ldots, \sigma(\mathbf{x}_n))$. The (local) variogram of the heteroscedastic spatial process is given by:

$$2\gamma_Y(\mathbf{x}, \mathbf{x}+\mathbf{u}) = (\sigma(\mathbf{x}) - \sigma(\mathbf{x}+\mathbf{u}))^2 + 2\sigma(\mathbf{x})\sigma(\mathbf{x}+\mathbf{u})\gamma(\mathbf{u}).$$

Under the general spatial model (2), a resampling approach will be proposed for approximation of (1). The new methodology extends the bootstrap procedure developed in [4] for the homoscedastic case, which is modified to adequately reproduce the variability of the data.

2 Main results

In what follows, we describe the suggested bootstrap procedure. Firstly, nonparametric estimators of the trend, variance and variogram functions must be obtained, which will be respectively denoted by $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\gamma}$. Starting by the spatial trend, its local linear estimator is given by:

$$\hat{\mu}(\mathbf{x}) = \mathbf{e}_1^t \left(\mathbf{X}_\mathbf{x}^t \mathbf{W}_\mathbf{x} \mathbf{X}_\mathbf{x}\right)^{-1} \mathbf{X}_\mathbf{x}^t \mathbf{W}_\mathbf{x} \mathbf{Y} = s_\mathbf{x}^t \mathbf{Y}, \tag{3}$$

where $\mathbf{e}_1 = (1, 0, ..., 0)^t \in \mathbb{R}^{d+1}$, $\mathbf{X}_\mathbf{x}$ is a matrix with i-th row equal to $(1, (\mathbf{x}_i - \mathbf{x})^t)$, $\mathbf{W}_\mathbf{x} = diag\left\{K_\mathbf{H}(\mathbf{x}_1 - \mathbf{x}), \ldots, K_\mathbf{H}(\mathbf{x}_n - \mathbf{x})\right\}$, $K_\mathbf{H}(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$, K is a d-dimensional kernel function and $\mathbf{H}$ represents the bandwidth matrix.

Then, the natural procedure to obtain the estimators $\hat{\sigma}$ and $\hat{\gamma}$ consists of first removing the trend and then estimating the variance and the variogram from the residuals $\mathbf{r} = \mathbf{Y} - \mathbf{SY}$, where $\mathbf{S}$ is the *smoother*

matrix, whose i-th row is equal to $\mathbf{s}_{\mathbf{x}_i}^t$. However, the direct use of the residuals tends to produce an underestimation of the small-scale variability of the process (e.g. [1], Section 3.4.3). Indeed:

$$Var(\mathbf{r}) = \boldsymbol{\Sigma} + \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^t - \boldsymbol{\Sigma}\mathbf{S}^t - \mathbf{S}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathbf{r}}.$$

Equivalently, the covariance matrix of the (unobserved) standardized residuals $\tilde{\varepsilon} = \mathbf{D}^{-1}\mathbf{r}$ is as follows:

$$Var(\tilde{\varepsilon}) = \mathbf{R} + \mathbf{B} = \boldsymbol{\Sigma}_{\tilde{\varepsilon}}, \tag{4}$$

with:

$$\mathbf{B} = \mathbf{D}^{-1}\left(\mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^t - \boldsymbol{\Sigma}\mathbf{S}^t - \mathbf{S}\boldsymbol{\Sigma}\right)\mathbf{D}^{-1}. \tag{5}$$

From (4), it is easy to see that:

$$Var\left(r_i/\sqrt{1+b_{ii}}\right) = \sigma^2(\mathbf{x}_i),$$

$$Var\left(\tilde{\varepsilon}(\mathbf{x}_i) - \tilde{\varepsilon}(\mathbf{x}_j)\right) = Var\left(\varepsilon(\mathbf{x}_i) - \varepsilon(\mathbf{x}_j)\right) + b_{ii} + b_{jj} - 2b_{ij},$$

where r_i is the i-th term of vector $\mathbf{r}$, b_{ij} is the (i,j)-th element of matrix $\mathbf{B}$ and $\tilde{\varepsilon}(\mathbf{x}_i) = r(\mathbf{x}_i)/\sigma(\mathbf{x}_i)$ is the i-th component of $\tilde{\varepsilon}$.

From these results, an iterative algorithm is designed for the joint estimation of the variance and the variogram. This method is similar to the one described in [3], although we propose using the "exact" bias matrix (5) instead of an approximation to it. The specific steps are summarized below:

1. Estimate the trend through (3), compute the residuals $\mathbf{r}$ and obtain a pilot (uncorrected) estimate of σ^2 by linear smoothing of $\left(\mathbf{x}_i, r_i^2\right)$.
2. Compute the estimated standardized residuals $\hat{\varepsilon} = \hat{\mathbf{D}}^{-1}\mathbf{r}$, with $\hat{\mathbf{D}} = diag(\hat{\sigma}(\mathbf{x}_1), \ldots, \hat{\sigma}(\mathbf{x}_n))$, and derive an estimator $\hat{\gamma}_{\hat{\varepsilon}}$ of the error semivariogram by linear smoothing of $\left(||\mathbf{x}_i - \mathbf{x}_j||, (\hat{\varepsilon}(\mathbf{x}_i) - \hat{\varepsilon}(\mathbf{x}_j))^2\right)$, where $\hat{\varepsilon}(\mathbf{x}_i)$ is the i-th component of $\hat{\varepsilon}$.
3. Obtain $\hat{\boldsymbol{\Sigma}}_{\hat{\varepsilon}}$ from $\hat{\gamma}_{\hat{\varepsilon}}$ and take $\hat{\mathbf{R}} = \hat{\boldsymbol{\Sigma}}_{\hat{\varepsilon}}$ as a pilot estimate of the correlation matrix.
4. Form $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{D}}\hat{\mathbf{R}}\hat{\mathbf{D}}$ and $\hat{\mathbf{B}} = \hat{\mathbf{D}}^{-1}\left(\mathbf{S}\hat{\boldsymbol{\Sigma}}\mathbf{S}^t - \hat{\boldsymbol{\Sigma}}\mathbf{S}^t - \mathbf{S}\hat{\boldsymbol{\Sigma}}\right)\hat{\mathbf{D}}^{-1}$.
5. Derive an updated estimate $\hat{\sigma}^2$ of the variance by linear smoothing of $\left(\mathbf{x}_i, r_i^2/(1+\hat{b}_{ii})\right)$ and take $\hat{\mathbf{D}} = diag(\hat{\sigma}(\mathbf{x}_1), \ldots, \hat{\sigma}(\mathbf{x}_n))$.
6. Recalculate $\hat{\varepsilon} = \hat{\mathbf{D}}^{-1}\mathbf{r}$ and approximate the error semivariogram by linear smoothing of $\left(||\mathbf{x}_i - \mathbf{x}_j||, (\hat{\varepsilon}(\mathbf{x}_i) - \hat{\varepsilon}(\mathbf{x}_j))^2 - \hat{b}_{ii} - \hat{b}_{jj} + 2\hat{b}_{ij}\right)$, where $\hat{b}_{ij}$ is the (i,j)-th element of matrix $\hat{\mathbf{B}}$.
7. Obtain a new estimate $\hat{\mathbf{R}}$ of the correlation matrix and repeat steps 4-7 up to convergence.

The previous estimates allow us to implement a bootstrap replicate of the heteroscedastic process Y, from the available data, and then use it for risk estimation, at a given location $\mathbf{x}_0 \in D$, as follows:

1. Use the previous algorithm to obtain both estimates $\hat{\mathbf{R}}$ (in the final step) and $\hat{\boldsymbol{\Sigma}}_{\hat{\varepsilon}}$ (in step 3), as well as their corresponding Cholesky factorizations $\hat{\mathbf{R}} = \mathbf{L}\mathbf{L}^t$ and $\hat{\boldsymbol{\Sigma}}_{\hat{\varepsilon}} = \mathbf{L}_{\hat{\varepsilon}}\mathbf{L}_{\hat{\varepsilon}}^t$.
2. Compute the "independent" variables $\mathbf{e} = \mathbf{L}_{\hat{\varepsilon}}^{-1}\hat{\varepsilon}$ and center them to derive an independent bootstrap sample of size n, denoted by $\mathbf{e}^*$.
3. Construct the bootstrap errors $\hat{\varepsilon}^* = (\hat{\varepsilon}^*(\mathbf{x}_1), \ldots, \hat{\varepsilon}^*(\mathbf{x}_n))^t$ by taking $\hat{\varepsilon}^* = \mathbf{L}\mathbf{e}^*$.

4. Derive the bootstrap replicas $\mathbf{Y}^* = (Y^*(\mathbf{x}_1), \cdots, Y^*(\mathbf{x}_n))^t$, with $Y^*(\mathbf{x}_i) = \hat{\mu}(\mathbf{x}_i) + \hat{\sigma}(\mathbf{x}_i)\hat{\varepsilon}^*(\mathbf{x}_i)$, for $i = 1, \ldots, n$.

5. Predict the value of $Y(\mathbf{x}_0)$ by $\hat{Y}^*(\mathbf{x}_0) = \hat{\mu}(\mathbf{x}_0) + \hat{\sigma}(\mathbf{x}_0)\hat{\varepsilon}^*(\mathbf{x}_0)$, with $\hat{\varepsilon}^*(\mathbf{x}_0)$ obtained through the application of the simple kriging approach on $\hat{\varepsilon}^*$.

By repeating this scheme a large number of times, an estimate of the target risk (1) at $\mathbf{x}_0$ is provided by the proportion of values $\hat{Y}^*(\mathbf{x}_0)$ exceeding the fixed threshold c.

Acknowledgments. The first author's work has been partially supported by the Universidad de las Fuerzas Armadas ESPE (Ecuador). The second author acknowledges financial support from the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G-01, Spain) and from MINECO grants MTM2014-52876-R and MTM2017-82724-R, all of them through the European Regional Development Fund (ERDF). The third author's work has been partially supported by the Spanish National Research and Development Program project [TEC2015-65353-R], by the ERDF and by the Xunta de Galicia (Spain) under project GRC2015-018 and under agreement for funding AtlantTIC (Atlantic Research Center for Information and Communication Technologies).

References

[1] Cressie, N. (1993). *Statistics for Spatial Data*. Wiley. New York.

[2] Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall. London.

[3] Fernández-Casal, R., Castillo-Páez, S. and García-Soidán, P. (2017). Nonparametric estimation of the small-scale variability of heteroscedastic spatial processes. *Spatial Statistics* **22**, 358–370.

[4] Fernández-Casal, R., Castillo-Páez, S. and Francisco-Fernández, M. (2018). Nonparametric geostatistical risk mapping. *Stochastic Environmental Research and Risk Assessment* **32**, 675–684.

[5] García-Soidán, P. and Menezes, R. (2017). Nonparametric construction of probability maps under local stationarity. *Environmetrics* **28**, e2438.

[6] Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press. Oxford.

[7] Krzysztofowicz, R. and Sigrest, A. A. (1997). Local climatic guidance for probabilistic quantitative precipitation forecasting. *Monthly Weather Review* **12**, 305–316.

[8] Li, W., Zhang, C., Dey, D. K. and Wang, S. (2010). Estimating threshold-exceeding probability maps of environmental variables with Markov chain random fields. *Stochastic Environmental Research and Risk Assessment* **24**, 1113–1126.

[9] Shapiro, A. and Botha, J. D. (1991). Variogram fitting with a general class of conditionally non-negative definite functions. *Computational Statistics and Data Analysis* **11**, 87–96.

[10] Tolosana-Delgado, R., Pawlowsky-Glahn, V. and Egozcue, J. J. (2008). Indicator kriging without order relation violations. *Mathematical Geosciences* **40**, 327–347.

Spatial analysis of crash data in the road network of the city of Valencia

Á. Briz[1,*], F. Martínez[2] and F. Montes[1]

[1] *Department of Statistics and Operational Research, University of Valencia, Spain; alvaro.briz@uv.es, francisco.montes@uv.es*
[2] *Oficina d'Estadística, Ajuntament de Valencia, Spain; fmartinezr@valencia.es*
[*] *Corresponding author*

Abstract. *In the last decade, research on traffic accidents is increasing and expanding with the design of a wide variety of methodologies and techniques. In order to properly capture the distribution of accidents within an area of interest, spatial models have become required, focusing on two main objectives: the detection of zones of high crash risk (hotspots) and the explicability of accidents incidence from a set of related variables of different nature. These include geometric features and traffic information of the roads being studied, but also socioeconomic and demographic issues.*
Most of the studies on traffic accidents have been developed over an areal region subject to a certain administrative division. However, recent studies have employed the specific network structure of roads of the region of interest, which are known as linear networks in the field of spatial statistics. These are like graphs made of edges and vertex that accurately represent the geographical space where accidents take place.
In this work, a collection of accidents registered by the Local Police Department of the city of Valencia (Spain) in the period 2005-2017 are projected to a linear network, which represents a zone of this city with more than 30 km of road structure. Furthermore, the linear network has been endowed with a direction according to the traffic flow. Several models including various road features as variables and different definitions of neighbourhoods have been analyzed and compared with explanatory and methodological objectives.

Keywords. *Spatial Statistics; Linear Networks; Crash Data.*

---

1 Introduction

In the last decade research on traffic accidents is increasing and expanding with the inclusion of a wide variety of statistical methodologies and techniques. The use of spatial lag models considering a rate of interest as the response variable has been successfully used since decades ago and applied to several kinds of geographic events, including traffic accidents [1]. However, many other models of different nature and scope have appeared recently to treat traffic accidents, some of which are now briefly summarized. In [2] traffic accident counts were modelled at the ward level employing non-spatial (negative binomial) and spatial methods based on Bayesian hierarchical models. In [3] the influence of road networks in the incidence of pedestrian-vehicles crashes was analyzed by developing a measure (integration) which reflects the accessibility of a node in the network, depending on its neighbourhood geometry. Finally, in [4] crash frequency at the county-level was explored with a spatial Bayesian model which included road

and traffic-related factors, but also demographic and socioeconomic information of the counties being studied. This work focused on the distinction of two types of exposure variables: population and average daily vehicle miles travelled per county.

2 Data

2.1 Accidents dataset

A total of 4548 traffic accidents registered by the Police Department of the city of Valencia (Spain) during the years 2005 to 2017 and located in the Eixample District of this city have been employed. According to their reported coordinates, these accidents were projected to a linear network representing the traffic roads of the Eixample District of Valencia (Figure 1 contains this linear network and a description of the counts observed at the road level).

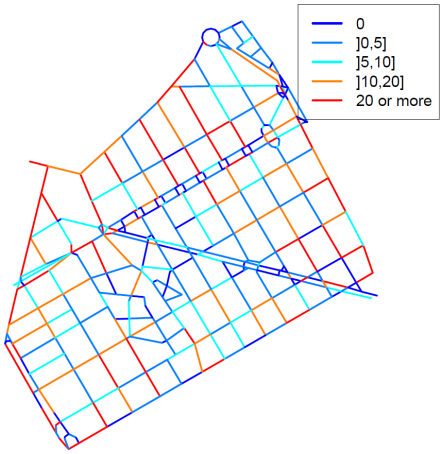


Figure 1: Accident counts at the road segment level for the Eixample District in the period 2005-2017.

2.2 Network structure and related variables

A linear network composed of 279 vertex and 444 edges, which represented a total length of 33.56 km was used for the analysis. Some parts of this network were previously simplified, including the slight modification of highly complex intersections and the removal of pedestrian streets. For the purpose of improving the analysis, the network was endowed with a direction according to the traffic of this district of Valencia as of the end of December of 2017.

Several factors that could influence in vehicle collisions, which are related to public infrastructures and road characteristics were considered at the edge level: parking lots, bus stops, traffic lights, number of lanes in the road, presence of a bus lane (binary), the road type (binary, main or not), and the number of roads that directly connect to the edge (number of neighbours). As the network of study represents a quite little and homogeneous population area, the inclusion of demographic or socioeconomic variables lacks of interest.

3 Methodology

3.1 Concept of neighbour

Given an edge, i, of the directed linear network, its neighbourhood, $N(i)$, can be defined in four different ways depending on whether the traffic flow information available is used. In the simplest way, if this information is not used, two edges i and j are neighbours if they share a vertex. However, the use of the traffic flow leads to the definition of three other types of neighbourhoods. First, the neighbourhood between i and j can be established if it is possible to travel from i to j or from j to i without passing through another edge of the network, indistinctly, which is denoted $N_{dir}(i)$. In addition, if a distinction is made between travelling from i to j, or vice versa, it is possible to define the neighbouring edges that allow you to reach i ($N^{in}_{dir}(i)$) of those that allow you to leave from i to another edge of the network ($N^{out}_{dir}(i)$) (see Figure 2 to see an example of all these neighbourhoods). Hereinafter these two types of neighbours are referred to as in-neighbours and out-neighbours, respectively.

The four ways of constructing neighbourhoods lead to the definition of four different adjacency matrices, which are named as $W, W_{dir}, W^{in}_{dir}, W^{out}_{dir}$, preserving the notation chosen for the neighbourhoods. For any of these matrices, their elements, w_{ij}, are called weights and $w_{ij} = 1$ if $j \in N(i)$, and 0 otherwise. The values of non-zero entries of these matrices are normalized to sum to 1 for every row. Note that both matrices W and W_{dir} are symmetric according to their definition, but W^{in}_{dir} and W^{out}_{dir} are generally not.

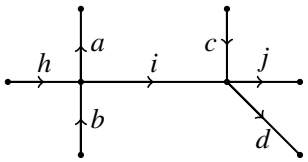


Figure 2: Examples of neighbourhoods in a directed linear network. The six edges that are contiguous to edge i allow the construction of the neighbourhoods $N(i) = \{a,b,c,d,h,j\}$, $N_{dir}(i) = \{b,d,h,j\}$, $N^{in}_{dir}(i) = \{b,h\}$ and $N^{out}_{dir}(i) = \{d,j\}$

3.2 Edge neighbourhood geometry

Geometric structure surrounding each edge of the network was studied. Given an edge of the network, the factors considered for each neighbouring edge were the neighbourhood type (in or out) and the angles formed between the edge and its neighbours. Edge length and the number of in and out neighbours were also included to better discriminate between edges. By using this information, a total of fifteen geometric-related variables were defined by combining the angles and lengths of the edges, the neighbourhood structure and the traffic flow. The k-means algorithm [5] was then applied to these fifteen variables, which allowed the composition of four clusters of 42, 130, 163 and 109 edges, respectively. Cluster 2 is mainly composed of medium-long edges that are part of a crossroad (intersection). Cluster 3 is formed with very short edges with a high proportion of sharp angles, which represent abrupt changes of direction in the directed network. Cluster 1 clearly presents the highest edge length and an high number of neighbours. Finally, Cluster 4 is made of short-medium length edges and a quite high connectivity with short-length edges if compared with Clusters 1 and 2.

3.3 Models tested

Three statistical models were tested and compared: spatial lag model (SL) [6] with two different definitions of the response variable and a basic non-spatial generalized linear model (GLM) with the accident counts at the edge level as the response. For the SL model the rate of accidents by road meter during the period of years considered was firstly used. Moreover, another continuous response variable was constructed by computing the equal-continuous kernel density estimate at the middle points of each of the 444 edges, according to the formula established in [7] choosing a bandwidth of 150 metres.

4 Results

The SL model based on the kernel density estimates clearly outperformed the other two in terms of model fitting, presenting a much lower AIC value. The use of the four different neighbourhoods did not affect the results as much as expected, but the W_{dir} neighbourhood matrix, considering both in and out neighbours, produced the best results. If the best fitting model (SL with kernel response) employing the W_{dir} matrix is selected for further discussion, it can be concluded that roads with only 2 lanes are a synonym of less traffic accidents. However, roads with five lanes and roads with two accesses (two in-neighbours) seem to produce more accidents. Finally, roads belonging to the aforementioned Cluster 1 are less dangerous than the ones that are part of the other three, despite the high presence of long roads with a high connectivity in this cluster.

Acknowledgments. We thank José Serrano, Chief of Valencia Local Police Department and the *Unitat d'Atestats i Seguretat en el Transport* of the city of Valencia.

References

[1] Levine, N., Kim, K. E., & Nitz, L. H. (1995). Spatial analysis of Honolulu motor vehicle crashes: I. Spatial patterns. *Accident Analysis & Prevention* **27**(5), 663–674.

[2] Quddus, M. A. (2008). Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. *Accident Analysis & Prevention* **40**(4), 1486–1497.

[3] Guo, Q., Xu, P., Pei, X., Wong, S. C., & Yao, D. (2017). The effect of road network patterns on pedestrian safety: A zone-based Bayesian spatial modeling approach. *Accident Analysis & Prevention* **99**, 114–124.

[4] Huang, H., Abdel-Aty, M., & Darwiche, A. (2010). County-level crash risk analysis in Florida: Bayesian spatial modeling. *Transportation Research Record: Journal of the Transportation Research Board* **(2148)**, 27–37.

[5] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**(1), 100–108.

[6] Anselin, L., & Bera, A. K. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. *Statistics Textbooks and Monographs* **155**, 237–290.

[7] McSwiggan, G., Baddeley, A., & Nair, G. (2017). Kernel density estimation on a linear network. *Scandinavian Journal of Statistics* **44**(2), 324–345.

Random permutation test in factorial models for spatial point patterns

Jonatan A. González [1*] and Jorge Mateu[2]

[1] *Department of Mathematics, Universitat Jaume I, Spain; jmonsalv@uji.es, mateu@uji.es*

Abstract. *We formulate statistical tests to check for interaction effects under the two-way ANOVA models when the observations are second-order descriptors of spatial point patterns. The data involved come from a metallurgy procedure related to flotation cells. In particular, we analyse the interaction effect between the frother concentration and the volumetric air flow factors in the spatial distribution of bubbles.*

Keywords. *Factorial models; Flotation bubbles; K-function; Point processes; Replicated point patterns.*

1 Introduction

In order to quantify the gas dispersion, well-known *gas dispersion properties* are defined within the flotation process [7, 3]. For instance, *gas holdup* (volumetric fraction of gas in a gas-slurry mix), *superficial gas velocity* (volumetric gas flow rate per cross-sectional area of cell), *bubble size distribution* (BSD, characterised by a statistical bubble diameter), and the derived parameter *bubble surface area flux* [5, 7].

Some of these gas dispersion properties can be measured from a snapshot of bubbles taken in some moment of the flotation experiment. The images of these bubbles are captured by a camera on top of a flotation cell. Each bubble represents a location with diameter or area of the bubble attached, and this can be considered a marked point pattern. We depict an example of such bubbles in Figure 1. The

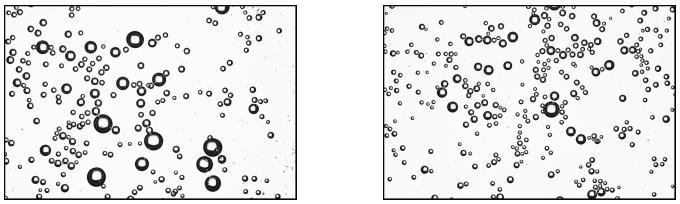


Figure 1: Two different sampled images of bubbles captured on top of a flotation cell. A tube, half inch in diameter, is immersed in the pulp, and the bubbles which are obtained from the foam are sampled.

characteristics of the flotation experiment are strongly dependent on some operating factors. A proper combination of gas rate and bubble size is required to provide a considerable gas holdup in the flotation pulp [6]. The volumetric air flow (L/min) and the specific frother concentration (ppm) are two factors

that could influence the physical properties [4]. Therefore, a relevant question is whether the combined action of these two factors enhances or inhibits the action of each other in the spatial descriptor. We develop a suitable statistical test by using the replication in order to analyse the interaction.

2 Methodology

Suppose that factor A has a levels and factor B has b levels. Each realisation or replicate contains all ab factorial combinations. Additionally let c be the fixed number of replicates in each cell. We have a functional descriptor sample $\{\hat{K}_{ijk}\}, i = 1,\ldots,a, j = 1,\ldots,b, k = 1,\ldots,c$.

For a two-way ANOVA, the mean $\mathcal{K}_{ij}(r)$ can be written as

$$\mathcal{K}_{ij}(r) = \mathcal{K}_0(r) + \tau_i(r) + \beta_j(r) + (\tau\beta)_{ij}(r), \qquad i = 1,\ldots,a, j = 1\ldots,b, r \in T, \tag{1}$$

where $\mathcal{K}_0(r)$ is the overall mean effect, $\tau_i(r)$ is the effect of the ith level of the *row factor A*, $\beta_j(r)$ is the effect of the jth level of *column factor B*, $(\tau\beta)_{ij}(r)$ is the effect of the interaction between $\tau_i(r)$ and $\beta_j(r)$. Both factors are assumed to be fixed, and the factor effects are defined as deviations from the overall mean. We assume balanced factorial designs, so we determine whether row and column factors interact, i.e, we test

$$\begin{aligned} &\mathcal{H}_0^I : (\tau\beta)_{ij}(r) = 0 \qquad \text{for all } i, j, \text{ and for } r \in T, \\ &\mathcal{H}_1^I : \text{at least one } (\tau\beta)_{ij}(r) \neq 0, \text{ for some } r \in T. \end{aligned} \tag{2}$$

We consider pooled estimators of cell weighted mean and covariance functions given by

$$\bar{K}_{ij\cdot}(r) = \frac{1}{\omega_{ij\cdot}} \sum_{k=1}^{c} \omega_{ijk} \hat{K}_{ijk}(r), \qquad i = 1,\ldots,a, j = 1,\ldots,b, \tag{3}$$

and

$$\hat{\gamma}_{ij}(r,s) = \frac{1}{c-1} \sum_{i=1}^{c} \left[\hat{K}_{ijk}(r) - \bar{K}_{ij\cdot}(r)\right] \left[\hat{K}_{ijk}(s) - \bar{K}_{ij\cdot}(s)\right], \tag{4}$$

where the number of points per pattern is denoted by n_{ijk}, where k is the individual within the ij cell (sample) and $i = 1,\ldots,a$ and $j = 1,\ldots,b$, and $\omega_{ij\cdot} = \sum_{k=1}^{c} n_{ijk}(n_{ijk}-1)$. As in the classical ANOVA two-way analysis, we define $\bar{K}_{i\cdot\cdot}, \bar{K}_{\cdot j\cdot}$ and $\bar{K}_{\cdots}$ as the corresponding row, column, and grand weighted average K-functions. Thus,

$$\begin{aligned} \bar{K}_{i\cdot\cdot}(r) &= \frac{1}{\omega_{i\cdot\cdot}} \sum_{j=1}^{b} \omega_{ij\cdot} \bar{K}_{ij\cdot}(r), \qquad \bar{K}_{\cdot j\cdot}(r) = \frac{1}{\omega_{\cdot j\cdot}} \sum_{i=1}^{a} \omega_{ij\cdot} \bar{K}_{ij\cdot}(r), \quad i = 1,\ldots,a, j = 1,\ldots,b, \\ \bar{K}_{\cdots}(r) &= \frac{1}{\omega_{\cdots}} \sum_{i=1}^{a} \sum_{j=1}^{b} \omega_{ij\cdot} \bar{K}_{ij\cdot}(r), \end{aligned} \tag{5}$$

where

$$\omega_{i\cdot\cdot} = \sum_{j=1}^{b} \omega_{ij\cdot}, \quad \omega_{\cdot j\cdot} = \sum_{i=1}^{a} \omega_{ij\cdot} \quad \text{and} \quad \omega_{\cdots} = \sum_{i=1}^{a} \sum_{j=1}^{b} \omega_{ij\cdot}. \tag{6}$$

From (5), the estimator of interaction effects is

$$\widehat{(\tau\beta)}_{ij}(r) = \bar{K}_{ij\cdot}(r) - \bar{K}_{i\cdot\cdot}(r) - \bar{K}_{\cdot j\cdot}(r) + \bar{K}_{\cdots}(r). \tag{7}$$

Let us consider some fixed $r \in T$ and let $\mathrm{SSI}(r)$ be the interaction-effect pointwise sum-of-squares and let $\mathrm{SSE}(r)$ denote the pointwise sum-of-squares due to errors. Following the classical balanced two-way ANOVA, we define

$$\begin{aligned} \mathrm{SSI}(r) &= \sum_{i=1}^{a}\sum_{j=1}^{b}[\bar{K}_{ij\cdot}(r) - \bar{K}_{i\cdot\cdot}(r) - \bar{K}_{\cdot j\cdot}(r) + \bar{K}_{\cdots}(r)]^2 = c\sum_{i=1}^{a}\sum_{j=1}^{b}\widehat{(\tau\beta)}_{ij}^{2}(r), \\ \mathrm{SSE}(r) &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left[\hat{K}_{ijk}(r) - \bar{K}_{ij\cdot}(r)\right]^2 = (c-1)\sum_{i=1}^{a}\sum_{j=1}^{b}\hat{\sigma}_{ij}^{2}(r). \end{aligned} \tag{8}$$

The corresponding Fisher-test type statistics for $\mathcal{H}_0^I$ is given by

$$F^I = \frac{\int_T \mathrm{SSI}(r)\mathrm{d}r/\left((a-1)(b-1)\right)}{\int_T \mathrm{SSE}(r)\mathrm{d}r/(ab(c-1))}. \tag{9}$$

2.1 Inference

Our interest focuses on testing the null hypothesis that K-functions are not altered by the combined effect of the factors even though the analytical form of the probability function of F^I remains unknown. Thus, we perform a pure randomisation test to permute exchangeable residuals across levels of the factors in order to obtain the approximate conditional distribution of F^I. We generate random samples as follows: in the first step, residual functions are defined as

$$\begin{aligned} \hat{\varepsilon}_{ijk}(r) = \quad & \sqrt{n_{ijk}(n_{ijk}-1)}\left[\hat{K}_{ijk}(r) - \bar{K}_{ij\cdot}(r)\right], \\ & i = 1,\ldots,a, j = 1,\ldots,b, k = 1,\ldots c, \end{aligned} \tag{10}$$

and

$$\begin{aligned} \hat{\varepsilon}^{\dagger}_{ijk}(r) = \quad & \sqrt{n_{ijk}(n_{ijk}-1)}\left[\hat{K}_{ijk}(r) - \hat{K}_{i\cdot\cdot}(r) - \hat{K}_{\cdot j\cdot}(r) + \hat{K}_{\cdots}(r)\right], \\ & i = 1,\ldots,a, j = 1,\ldots,b, k = 1,\ldots c. \end{aligned} \tag{11}$$

Under the null hypothesis, $\hat{\varepsilon}_{ijk}(r)$ and $\hat{\varepsilon}^{\dagger}_{ijk}(r)$ are approximately exchangeable quantities since the sampling variance of each $K_{ijk}(r)$ is proportional to n_{ijk}^{-1}. We analyse a set of simulations generated by varying parameters as the intensity in the standard Poisson case. Additionally we simulate patterns with different clustering parameters. The simulation study shows that permutation of residuals reproduce the distribution of the test statistic, presenting approximate uniformity p-values in most cases as well as good behaviour concerning the power.

2.2 Data analysis

The test based on Fisher-type statistic was carried out using the integration interval $T = (0, r_0]$, whose upper integration bound is given by $r_0 = 1.57$mm. We implemented the tests with 500000 random permutations. For the interaction effect, we have a significant p-value associated with the residuals $\hat{p}_{\hat{\varepsilon}_{ijk}} = 0.0275$, analogously $\hat{p}_{\hat{\varepsilon}^{\dagger}_{ijk}} = 0.0153$.

The point patterns exhibit significant differences up to small distances (see Figure 2). We conclude that the bubble patterns do not support hypothesis of zero interaction. The K-functions indicate that the effects of the frother concentration differ for the three levels of the volumetric airflow rate.

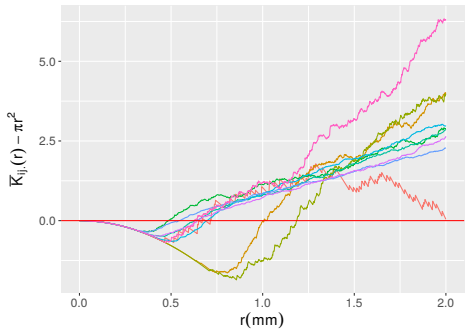


Figure 2: Pooled mean values of $\hat{K}_{ijk}(r)$ estimated on 9 cells of the flotation experiment. The red line represents the complete spatial randomness.

Acknowledgments. The authors are partially funded by grants MTM2016-78917-R and P1-1B2015-40.

References

[1] Baddeley, A., Rubak, E., Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R.* CRC Press, Boca Raton, Florida.

[2] Diggle, P.J., Lange, N., Benes, F. (1991). Analysis of variance for replicated spatial point patterns in clinical neuroanatomy. *Journal of the American Statistical Association* **86**, 618-625.

[3] Gómez, C.O., Finch, J.A. (2007). Gas dispersion measurements in flotation cells. *International Journal of Mineral Processing* **84**, 51-58.

[4] Gómez C.O., Mesías J., Álvarez J.(2016). Bubble Surface Area Flux and Performance in laboratory flotation testing. *XXVIII International Mineral Processing Congress Proceedings.* Canadian Institute of Mining, Metallurgy and Petroleum.

[5] Gorain, B., Franzidis, J., Manlapig, E. (1999). The empirical prediction of bubble surface area flux in mechanical flotation cells from cell design and operating data. *Minerals Engineering* **12**, 309-322.

[6] Miskovic, S. Luttrell, G. (2009). Column and non-traditional flotation. *Recent Advances in Mineral Processing Plant Design.*.

[7] Nesset, J.E., Hernandez-Aguilar, J.R., Acuna, C., Gomez, C.O., Finch, J.A. (2006). Some gas dispersion characteristics of mechanical flotation machines. *Minerals Engineering* **19**, 807-815.

[8] Randall, E.W., Goodall, C.M., Fairlamb, P.M., Dold, P. L., O'Connor, C. T. (1989). A method for measuring the sizes of bubbles in two- and three-phase systems. *Journal of Physics E: Scientific Instruments* **22**, 827-833.

[9] Schlather, M., Ribeiro, P.J., Diggle, P.J. (2004). Detecting dependence between marks and locations of marked point processes. *Journal of the Royal Statistical Society, Series B* **66**, 79-93.

[10] Schwarz, S., Alexander, D. (2006). Gas dispersion measurements in industrial flotation cells. *Minerals Engineering* **19**, 554-560.

[11] Wilson, H.E., 1998. *Statistical analysis of replicated spatial point patterns*. Ph.D. thesis, University of Lancaster, United Kingdom.

Bootstrap bandwidth selection for kernel estimation of the pair correlation function in inhomogeneous spatial point processes

I. Fuentes-Santos[1,*], W. González-Manteiga[2,*] and J. Mateu[1]

[1] *Marine Research Institute, Spanish National Research Council, Vigo (Spain); isabel.fuentes@usc.es*
[2] *Department of Statistics, Mathematical Analysis and Optimization. University of Santiago de Compostela, (Spain); wenceslao.gonzalez@usc.es*
[3] *Department of Mathematics, University Jaume I, Castellón (Spain); mateu@mat.uji.es*
[*] *Corresponding author*

Abstract. *The pair correlation function can be considered one of the most informative second-order characteristic of spatial point processes. Nonparametric estimators of the pair correlation function are useful tools to identify the type, aggregation or inhibition, and strength of spatial interaction in observed spatial point patterns. Kernel smoothing is the most popular nonparametric estimator of the pair correlation function for both homogeneous and inhomogeneous point processes. The performance of any kernel estimator depends on the bandwidth parameter. Several procedures, such as cross-validation, semiparametric bootstrap or an adaptive plug-in rule, have been proposed for bandwidth selection in the stationary framework. To our knowledge, least-squares cross validation is the only data-driven bandwidth selector available for the inhomogenous case. This work analyzes the asymptotic properties of the kernel estimator of the pair correlation function for second-order intensity reweighted stationary (SOIRS) point processes. We propose a nonparametric bootstrap to estimate the asymptotic mean square error (AMISE), and develop a bandwidth selector based on the minimization of the bootstrap AMISE. We compare the performance of our proposal with the least-squares bandwidth selector in a simulation study, and through its application to the second-order analysis of wildfires patterns in Galicia (NW Spain).*

Keywords. *AMISE; Data-driven; Nonparametric bootstrap; Second-order characteristics; Wildfires*

Nonparametric approximation of conditional risk in non-stationary geostatistical processes

Fernández-Casal, R.[1,*], Castillo-Páez, S.[2] and Francisco-Fernández, M.[1]

[1] *Departamento de Matemáticas, Facultad de Informática, Universidade da Coruña, 15071, A Coruña, Spain; ruben.fcasal@udc.es, mariofr@udc.es*
[2] *Departamento de Ciencias Exactas, Universidad de las Fuerzas Armadas, Sangolquí, Ecuador; sacastillo@espe.edu.ec*
[*] *Corresponding author*

Abstract. *In this work, a nonparametric procedure to approximate the conditional probability that a regionalized variable exceeds a certain threshold value is proposed. The method consists of a bootstrap algorithm that combines conditional simulation techniques with nonparametric estimations of the trend and the variogram of the spatial process. For the local linear estimation of the mean, a bandwidth selection method that takes the spatial dependence into account is used. The variogram is approximated by a flexible estimator based on the residuals, previously correcting its bias due to the estimation of the trend. The proposed method allows obtaining estimates of the exceedance risk in non-observed spatial locations, and its behavior will be analyzed through simulation studies and with the application to a real data set.*

Keywords. *Conditional simulation; Local linear estimation; Bootstrap.*

---

1 Introduction

An important tool for the analysis of environmental problems is the construction of risk maps. These maps provide estimates of the probability that a given study variable (for example, pollutants, precipitation levels, etc.) exceeds certain permissible values.

Assuming the spatial process $\{Y(\mathbf{x}), \mathbf{x} \in D \subset \mathbb{R}^d\}$, our interest focuses on the estimation of the conditional probability $r_c(\mathbf{x}) = P(Y(\mathbf{x}) \geq c|\mathbf{Y})$, where $\mathbf{Y} = (Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n))^t$ denotes the observation vector and c is a threshold value.

Additionally, we will suppose that the process is not stationary in the mean, that is,

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + \varepsilon(\mathbf{x}), \tag{1}$$

where $\mu(\cdot)$ is the trend function, and the error term ε, representing the spatial dependence, is a second order stationary process with zero mean and covariogram $C(\mathbf{u}) = Cov(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{x}+\mathbf{u}))$, with $\mathbf{u} \in D$. However, in practice it is preferred to estimate the small-scale variability through the variogram $\gamma(\mathbf{u}) = C(\mathbf{0}) - C(\mathbf{u})$.

The geostatistical techniques commonly used to approximate this probability range from traditional methods, such as indicator kriging or disjunctive kriging (see, e.g. [6] y [10], respectively), to more recent procedures, such us those based on analysis of compositional data (see, e.g. [9]). However, these methods usually assume parametric models and, therefore, they can present misspecification problems. From the nonparametric point of view, Fernández-Casal *et al* [4] proposed a bootstrap method to estimate the unconditional probability $P(Y(\mathbf{x}) \geq c)$, using local linear estimates of the trend and the variogram, jointly with a procedure to correct the bias due to the use of residuals.

In the present work, we will extend the previous method using conditional simulation techniques, so that the behavior of the bootstrap replicas generated by this procedure match with the sampling behavior observed from the vector $\mathbf{Y}$.

2 Nonparametric estimation and unconditional bootstrap

Assuming model (1), the local linear estimator of the trend can be written as (see, e.g. [8]):

$$\hat{\mu}_{\mathbf{H}}(\mathbf{x}) = \mathbf{e}_1^t \left(\mathbf{X}_\mathbf{x}^t \mathbf{W}_\mathbf{x} \mathbf{X}_\mathbf{x}\right)^{-1} \mathbf{X}_\mathbf{x}^t \mathbf{W}_\mathbf{x} \mathbf{Y} \equiv s_\mathbf{x}^t \mathbf{Y}, \tag{2}$$

where $\mathbf{e}_1 = (1,0,\ldots,0)$, $\mathbf{X}_\mathbf{x}$ is a matrix with ith row equal to $(1,(\mathbf{x}_i - \mathbf{x})^t)$,

$$\mathbf{W}_\mathbf{x} = \mathtt{diag}\left\{K_\mathbf{H}(\mathbf{x}_1 - \mathbf{x}), ..., K_\mathbf{H}(\mathbf{x}_n - \mathbf{x})\right\},$$

with $K_\mathbf{H}(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$, K being a kernel function, and $\mathbf{H}$ a $d \times d$ matrix bandwidth. To select this bandwidth under dependence, the criterion proposed in [5] is recommended.

In a similar way, the local linear estimation of the variogram $\hat{\gamma}(\mathbf{u})$ can be obtained by applying (2) using the residuals $\mathbf{r} = \mathbf{Y} - \hat{\mu}_\mathbf{H}(\mathbf{x})$. However, this procedure introduces biases in this estimation, underestimating the variability of the spatial process (see e.g. [2], Section 3.4.3). Simply note that:

$$Var(\mathbf{r}) = \boldsymbol{\Sigma} + \boldsymbol{S}\boldsymbol{\Sigma}\boldsymbol{S}^t - \boldsymbol{\Sigma}\boldsymbol{S}^t - \boldsymbol{S}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\mathbf{r}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the errors and $\boldsymbol{S}$ is the $n \times n$ matix whose $i-$row is equal to $s_\mathbf{x}^t$. To reduce this effect, a nonparametric procedure similar to that proposed in [3] can be used, in order to obtain a corrected local linear estimator of the variogram $\hat{\gamma}(\cdot)$.

Based on the above, Fernández-Casal *et al* [4] proposed a bootstrap procedure to estimate the unconditional risk at a non-observed location $\mathbf{x}_\alpha$. In their approach, the estimated unconditional probability is obtained from kriging predictions, whose values, in practice, smooth the true spatial fluctuation of the data (see, e.g. [7]). As this could affect the estimation of the exceedance probability, we propose the following algorithm to generate unconditional bootstrap replicas $Y_{NS}^*(\mathbf{x}_\alpha)$ at the estimation locations $\{\mathbf{x}_\alpha : \alpha = 1,\ldots,n_0\}$:

1. Compute $\hat{\mu}_\mathbf{H}(\mathbf{x})$ and the corresponding residuals $\mathbf{r}$ to obtain $\hat{\gamma}_\mathbf{r}(\cdot)$ and its corrected version $\hat{\gamma}(\cdot)$, following [3].
2. Form $\hat{\boldsymbol{\Sigma}}_\mathbf{r}$ from $\hat{\gamma}_\mathbf{r}(\cdot)$, and find the matrix $\mathbf{L}$ such that $\hat{\boldsymbol{\Sigma}}_\mathbf{r} = \mathbf{L}_\mathbf{r}\mathbf{L}_\mathbf{r}^t$, using Cholesky decomposition.
3. Compute the "uncorrelated" residuals $\mathbf{e} = (e_1, e_2, \ldots, e_n)^t = \mathbf{L}_\mathbf{r}^{-1}\mathbf{r}$ and center them.

4. Obtain independent bootstrap samples of size n_0 from $\mathbf{e}$, denoted by $\mathbf{e}^* = \left(e_1^*, e_2^*, \ldots, e_{n_0}^*\right)^t$.
5. Form the covariance matrix $\hat{\boldsymbol{\Sigma}}_\alpha$ corresponding to the estimation locations $\mathbf{x}_\alpha$ using $\hat{\gamma}(\cdot)$, and compute $\mathbf{L}_\alpha$ such that $\hat{\boldsymbol{\Sigma}}_\alpha = \mathbf{L}_\alpha \mathbf{L}_\alpha^t$.
6. Compute the unconditional bootstrap errors $\boldsymbol{\varepsilon}_{NS}^* = (\varepsilon_{NS}^*(\mathbf{x}_1), \ldots, \varepsilon_{NS}^*(\mathbf{x}_{n_0}))^t$, such that $\boldsymbol{\varepsilon}_{NS}^* = \mathbf{L}_\alpha \mathbf{e}^*$.
7. Obtain the unconditional bootstrap replicas $Y_{NS}^*(\mathbf{x}_\alpha) = \hat{\mu}_{\mathbf{H}}(\mathbf{x}_\alpha) + \varepsilon_{NS}^*(\mathbf{x}_\alpha),\ \alpha = 1, \ldots, n_0$.

The latter algorithm uses non-conditional simulation techniques based on Cholesky's decomposition. This would allow us to assume that the bootstrap replicas have their mean and variance-covariance matrix equal to those of the spatial process $Y(\cdot)$, under the assumption of unbiasedness of the corresponding estimators (see, e.g. [2], Section 3.6.1.). However, as the behavior of these replicas does not necessarily coincide with the observed values at the sample locations (see, e.g. [1], Section 7.3.1), this algorithm should not be used for the estimation of the conditional risk.

3 Conditional bootstrap algorithm

For the sake of simplicity, let us assume that the trend is known in (1) and the true errors $\boldsymbol{\varepsilon} = (\varepsilon(\mathbf{x}_1), \ldots, \varepsilon(\mathbf{x}_n))$ are observed. The principle of the conditional simulation of the error at a location $\mathbf{x}_\alpha$ (see, e.g. [7]), starts from the trivial decomposition:

$$\varepsilon(\mathbf{x}_\alpha) = \hat{\varepsilon}(\mathbf{x}_\alpha) + [\varepsilon(\mathbf{x}_\alpha) - \hat{\varepsilon}(\mathbf{x}_\alpha)], \tag{3}$$

where $\hat{\varepsilon}(\mathbf{x}_\alpha)$ is the simple kriging prediction at $\mathbf{x}_\alpha$ computed from $\boldsymbol{\varepsilon}$. The idea is to substitute the unknown kriging error (the second term on the right), by a simulation of this error obtained from a non-conditional simulation $\varepsilon_{NS}(\mathbf{x})$ of the process. Then, a conditional simulation of the process is:

$$\varepsilon_{CS}(\mathbf{x}_\alpha) = \hat{\varepsilon}(\mathbf{x}_\alpha) + [\varepsilon_{NS}(\mathbf{x}_\alpha) - \hat{\varepsilon}_{NS}(\mathbf{x}_\alpha)]. \tag{4}$$

where $\hat{\varepsilon}_{NS}(\mathbf{x}_\alpha)$ is the kriging prediction at $\mathbf{x}_\alpha$ obtained from the unconditional simulations $\varepsilon_{NS}(\mathbf{x}_i)$ at the sample locations. Proceeding in this way, it is easy to verify that $\varepsilon_{CS}(\mathbf{x}_i) = \varepsilon(\mathbf{x}_i)$ and, in the case of simple kriging, $Cov(\varepsilon_{CS}(\mathbf{x}), \varepsilon_{CS}(\mathbf{x}+\mathbf{u})) = C(\mathbf{u})$ (see e.g. [1], Section 7.3.1). These properties guarantee that the simulations reproduce the behavior of the observed data, keeping the dependence structure of the spatial process.

Taking into account the previous results and that the trend $\hat{\mu}_{\mathbf{H}}$ of the bootstrap replicates is known, the proposed bootstrap algorithm to estimate the conditional risk is as follows:

1. Use the unconditional bootstrap algorithm described in previous section to (jointly) generate $\varepsilon_{NS}^*(\mathbf{x}_\alpha)$, $\alpha = 1, \ldots, n_0$ and $\varepsilon_{NS}^*(\mathbf{x}_i)$, $i = 1, \ldots, n$.
2. Compute the simple kriging predictions $\hat{\varepsilon}(\mathbf{x}_\alpha)$ and $\hat{\varepsilon}_{NS}^*(\mathbf{x}_\alpha)$ from the observed residuals $\mathbf{r}$ and from the bootstrap errors $\varepsilon_{NS}^*(\mathbf{x}_i)$, respectively.
3. Obtain the conditional bootstrap errors $\varepsilon_{CS}^*(\mathbf{x}_\alpha) = \hat{\varepsilon}(\mathbf{x}_\alpha) + [\varepsilon_{NS}^*(\mathbf{x}_\alpha) - \hat{\varepsilon}_{NS}^*(\mathbf{x}_\alpha)]$.
4. Compute the conditional bootstrap replicas $Y_{CS}^*(\mathbf{x}_\alpha) = \hat{\mu}_{\mathbf{H}}(\mathbf{x}_\alpha) + \varepsilon_{CS}^*(\mathbf{x}_\alpha)$.

5. Repeat steps 1 to 4 a large number of times B to obtain $Y_{CS}^{*(1)}(\mathbf{x}_\alpha), \ldots, Y_{CS}^{*(B)}(\mathbf{x}_\alpha)$.

6. Obtain $\hat{r}_c(\mathbf{x}_\alpha) = \frac{1}{B} \sum_{j=1}^{B} I\left(Y_{CS}^{*(j)}(\mathbf{x}_\alpha) \geq c\right)$.

References

[1] Chilès, J., and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. Wiley & Sons, New York.

[2] Cressie, N. (1993). *Statistics for Spatial Data*. Wiley & Sons, New York.

[3] Fernández-Casal, R. and Francisco-Fernández, M. (2014). Nonparametric bias-corrected variogram estimation under non-constant trend. *Stochastic Environmental Research and Risk Assessment* **28**, 1247–1259.

[4] Fernández-Casal, R., Castillo-Páez, S., y Francisco-Fernández, M. (2017a). Nonparametric geostatistical risk mapping. *Stochastic Environmental Research and Risk Assessment.* 32(3), 675-684

[5] Francisco-Fernández, M. and Opsomer, J. D. (2005). Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. *The Canadian Journal of Statistics* **33**, 279–295.

[6] Goovaerts, P., Webster, R. and Dubois, P. (1997). Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. *Environmental and Ecological Statistics* **4**, 31–48.

[7] Journel, A.G. (1974). Geostatistics for conditional simulation of Ore Bodies. *Economic Geology* **69**, 673–687.

[8] Opsomer, J. D., Wang, Y. and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science* **16**, 134–153.

[9] Tolosana-Delgado, R., Pawlowsky-Glahn, V., y Egozcue, J.-J. (2008). Indicator kriging without order relation violations. *Math Geosci.*, 40(3):327-347

[10] Webster, R. and Oliver, M.A. (1989). Optimal interpolation and isarithmic mapping of soil properties. VI. Disjunctive kriging and mapping the conditional probability. *Journal of Soil Science* **40**, 497–512.

Prevalence of obesity in Mexico: model for input values

Gutiérrez-Prieto, Ángel[1], Díaz-Avalos, Carlos[2] and Mejía-Domínguez, Nancy R.[3]*

[1] *Master in Mathematics, Institute in Applied Mathematics and Systems, Autonomus National University of Mexico, Mexico; angel.hipercubo@gmail.com*
[2] *Research Institute in Applied Mathematics and Systems, Autonomus National University of Mexico, Mexico; carlos@sigma.iimas.unam.mx*
[3] *Red de Apoyo a la Investigación. Coordinación de la Investigación Científica, Universidad Nacional Autónoma de México, Mexico; nmejia@cic.unam.mx*
**Corresponding author*

Abstract. *Obesity has become a health problem worldwide. According to the WHO, the tren in obesity in México is one of the highest worldwide. The association of obesity as a risk factor to hipertensive diseases and diabetes makes it necessary to analise the spatial distribution of its prevalence in the country. Such analyses provide information necessary to plan the level of health care that will be needed in the short and mid term. In this work we present the results of a preliminary analysis using the data provided by the 2012 National Health Survey. We fit a hierarchical poisson model to asses the relative prevalence at a municipality level. Our results show that the prevalence does not show any clear trend, an that there exist several hot spots associated to very particular municipalities.*

Keywords. *Non transmisible disease, obesity, spatial prevalence.*

---

1 Introduction

The prevalence of non-transmisible diseases is rising rapidly throughout the world, and it is expected to be the leading cause of mortality in Latin America [8]. The increase in the number of cases of obesity is a worldwide phenomenon and Mexico is not the exception. Moreover, annual prevalence rate of obesity in Mexico increased among adults by approximately 2 percent per year between 1988 and 2006, the largest increase documented worldwide [2]. Obesity is the main modifiable risk factor for the development of chronic non-communicable diseases, such as diabetes mellitus and cardiovascular diseases [7]. Also, diabetes mellitus and cardiovascular diseases are the two main causes of general mortality in Mexico. For type 2 diabetes a complex gene-environmental interaction for which several risk factors, such as age, sex, obesity and hypertension, are well documented [6]. In Mexico, near 11.7 million Mexicans are expected to have diabetes by the year 2025 [5]. Despite this scenario, there is a lack of spatial studies examining the association of diabetes or cardiovascular diseases with the spatial distribution of obesity or others socio-economic indicators. Although non-transmisible diseases cannot be characterized by an infectious agent, the observed spatial pattern of incidence (new events) or prevalent cases could provide information on the underlying mechanisms of the disease [1]. In this context, we asses this relationship using Bayesian spatial models at municipality level in Mexico to obtain maps of the spatial prevalence of obesity. The results shown here corresp

2 Methods

2.1 Data

We used the data from National Health Survey to Mexico (2012) and socio-economic indicators from National Institute of Geographic and Statistics form the 2010 census.

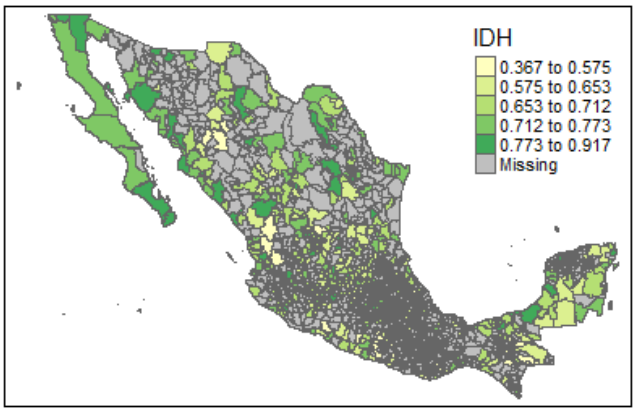


Figure 1: Human index of development for municipalities in Mexico

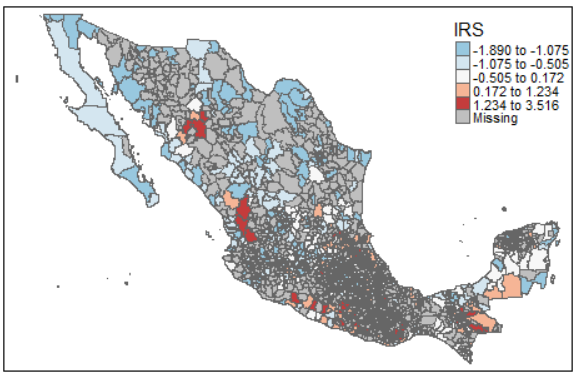


Figure 2: Social of lag index for municipalities in Mexico

2.2 Statistical models

For mapping the risk of obesity let y_i the observed number of people with obesity in the municipality i and let x_i the log relative risk in the zone i. We can assume that y_i given x_i are independent Poisson

variates due obesity is not a contagious disease. So,

$$y_i|x_i \sim Pois(c_i e^{x_i})$$

where c_i is the expected number of people with obesity in the municipality i with constant risk, as it has been modeled for other diseases [3]. The aim of this study is to model x_i considering the spatial dependence and socioeconomic factors as Human Ddevelopment Index (IDH) (figure **??**) and Social Gap Index (IRS) (Figure **??**) by municipality. We propose the standard model:

$$x_i = z_i^t \beta + u_i + v_i \text{ for } i \in \{1,...,n\} \quad (1)$$

where:
1.- $x_i = log(r_i)$ and r_i is relative risk for i municipality
2.- z_i is the associated vector of covariates for the i municipality
3.- β is the associated vector of coefficients to covariates.
4.- u_i represents variables that would display spatial structure.
5.- v_i is the effect of subjacents variables without structure. To estimate the parameters in (1) we use Bayesian approach based on MCMC methods.

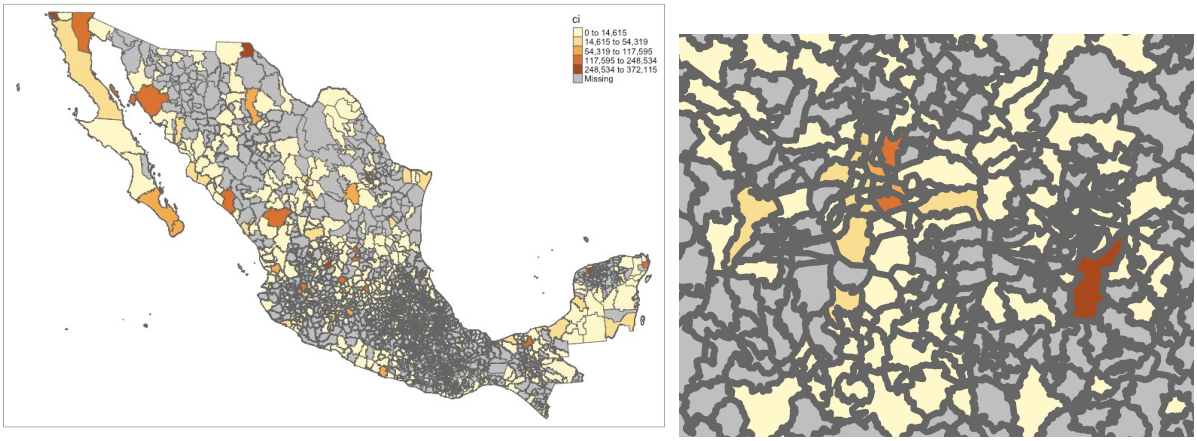


Figure 3: Estimated relative prevalence of obesity (X10) in the mexican municipalities (left) and a zoom to show the relative prevalence of obesity (X10) ina an area around México City.

3 Results and discussion

Figure 3 shows the results of the posterior mean of the relative prevalence of obesity in mexican municipalities and a zoom for the area around México City. The spatial distribution shoes the presence of several hot spots, where the relative prevalence of obesity is 24 fold time higher with respect to the national average. Those hot spots do not form a cluster and are scattered over different states. They correspond to rural municipalities with high social lag and IDH and high proportion of migrant workers in the United States, except for the municipality of Tijuana, in the upper left extreme in the map, which is a large city with a high proportion of people living in poverty. Tijuana has a very high proportion of migratory workers and also high number of people that has been deported back to México. Results about other factors associated to the spatial variability shown in figure 3 will be discussed.

References

[1] Banerjee, S., Carlin B.P., Gelfand, A.E. (2004) *Hierarchical Modelling and Analysis for Spatial Data*. Chapman and Hall, NY.

[2] Barquera S, Tovar-Guzman V, Campos-Nonato I, Gonzalez-Villalpando C, Rivera-Dommarco J (2003) Geography of diabetes mellitus mortality in Mexico: an epidemiologic transition analysis. *Arch Med Res 2003* **34**(5), 407–414.

[3] Gómez-Rubio, V., López-Quíles, M. (2010) Writing technical reports and papers. *The American Statistician* **36**, 326–329.

[4] Goovaerts, P. (2006). Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. *International Journal of Health Geographics* **5**, 1–31.

[5] King, H., Aubert, R., Herman, W. (1998). Global burden of diabetes, 1995-2025. *Diabetes Care* **21** 1414–1431.

[6] Noble, D., Mathur, R., Dent, T., Meads, C., Greenhalgh, T. (2011). Risk models and scores for type 2 diabetes: systematic review. *BMJ* **343** d7163.

[7] Rull, J.A., Águilar-Salinas, C.A., Rojas, R., Rios-Torres, J.M., Gómez-Pérez, F.J., Olaiz, G. (2005). Epidemiology of type 2 diabetes in Mexico. *Archives of Medical Research* **36** 188–196.

[8] Shaw, J.E., Sicree, R.A., Zimmet, P.Z. (2010). Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res. Clin. Pract* **87** 4–14.

Functional regression with spatially correlated errors

Johann Ospína-Galindez[1], Ramón Giraldo[2,*] and Mercedes Andrade-Bejarano[1]

[1] *School of Statistics, Universidad del Valle, Cali, Colombia; johann.alexis.ospina@gmail.com, mercedes.andrade@correounivalle.edu.co*
[2] *Statistics Department, Universidad Nacional de Colombia, Bogotá, Colombia; rgiraldoh@unal.edu.co*
[*] *Corresponding author*

Abstract. *We show an extension of the functional regression model to the case of spatially correlated errors. The estimation of the parameters is obtained by feasible generalized least squares. Functional geostatistics and particularly the trace-variogram function is proposed as a method for estimating the spatial dependence*

Keywords. *Feasible Least squares; Functional regression; Functional geostatistics; Trace-variogram*

1 Introduction

We propose a functional regression model to relate two functional variables (response and covariate) observed in sites (stations) of region. Data at each station are previously smoothed by using basis functions. This methodology allows carrying out spatial estimation of a response curve given the information of a functional covariate obtained at the same station. We consider a functional concurrent model [4] with spatially correlated errors. Functional geostatistics and in particular the trace-variogram function [2] is used to estimate the spatial dependence of errors curves [1]. The trace-variaogram parameters are estimated by the method-of-moments. Feasible generalized least squares is considered as estimation method of the regression parameters. Next we describe the essentials of the methodology proposed. We also show an application with rainfall curves [5]. Some conclusions are given at the end.

2 Functional regression model with spatial dependence

Suppose you have a collection of spatially indexed functional variables $(X_i(t), Y_i(t)), i = 1, \ldots, n$, with $(1, \cdots, n)$ a n-tuple of sites $\in \mathcal{D} \subset \mathbb{R}^2$ and we want to study the relationship between these ones by considering the concurrent model [4]

$$\begin{pmatrix} Y_1(t) \\ \vdots \\ Y_n(t) \end{pmatrix} = \begin{pmatrix} 1 & X_1(t) \\ \vdots & \vdots \\ 1 & X_n(t) \end{pmatrix} \begin{pmatrix} \beta_0(t) \\ \beta_1(t) \end{pmatrix} + \begin{pmatrix} \varepsilon_1(t) \\ \vdots \\ \varepsilon_n(t) \end{pmatrix}, \tag{1}$$

where $\beta_0(t)$ and $\beta_1(t)$ are the parameters of interest and $\varepsilon_i(t), i = 1,\ldots,n$ are n spatially correlated random errors. The problem considered here is the estimation of these parameters considering the spatial dependence of the model. It is assumed that both functional variables and parameters in model (1) can be expressed in terms of K basis functions, $\theta_1(t),\ldots,\theta_K(t)$, by

$$X_i(t) = \sum_{j=1}^{K} x_{ij}\theta_j(t) = \mathbf{x}_i^T\boldsymbol{\theta}(t) \tag{2}$$

$$\beta_0(t) = \sum_{j=1}^{K} b_{0j}\theta_j(t) = \boldsymbol{\theta}^T(t)\mathbf{b}_0 \tag{3}$$

$$\beta_1(t) = \sum_{j=1}^{K} b_{1j}\theta_j(t) = \boldsymbol{\theta}^T(t)\mathbf{b}_1. \tag{4}$$

With the representations in (2) to (4), the model (1) can be written, in matrix notation, as

$$\mathbf{Y}(t) = \mathbf{X}(t)\boldsymbol{\Theta}(t)\mathbf{b} + \boldsymbol{\epsilon}(t) \tag{5}$$

The estimation parameters $\beta_0(t)$ and $\beta_1(t)$ in 1 is obtained through the estimation of $\mathbf{b} = (\mathbf{b}_0, \mathbf{b}_1)^T$ in 5 by using generalized least squares. We minimize respect to $\mathbf{b}$

$$SSE(\mathbf{b}) = \int \left[\mathbf{Y}(t) - \mathbf{X}(t)\mathbf{b}\right]^T \Omega^{-1} \left[\mathbf{Y}(t) - \mathbf{X}(t)\mathbf{b}\right], \tag{6}$$

where Ω is the matrix of variances and covariances $n \times n$ of the vector $\boldsymbol{\epsilon}(t)$, with elements $\Omega_{ij} = Cov(\varepsilon_i(t), \varepsilon_j(t))$. Differentiating respect to $\mathbf{b}$ in (6) we have

$$\hat{\mathbf{b}} = \left[\int \mathbf{X}^T(t)\Omega^{-1}\mathbf{X}(t)dt\right]^{-1} \left[\int \mathbf{X}^T(t)\Omega^{-1}\mathbf{Y}(t)dt\right]. \tag{7}$$

In practice the covariance matrix of the errors Ω is unknown and consequently we cannot calculate $\hat{\mathbf{b}}$ in equation (7). The alternative that we consider is to use a feasible generalized least squares estimator (FLSE). First a functional concurrent regression model assuming independence is estimated. Then the residuals and the trace-variogram function [2, 3] are used to estimate the matrix Ω. Posteriorly we estimate the parameters by

$$\hat{\mathbf{b}} = \left[\int \mathbf{X}^T(t)\hat{\Omega}^{-1}\mathbf{X}(t)dt\right]^{-1} \left[\int \mathbf{X}^T(t)\hat{\Omega}^{-1}\mathbf{Y}(t)dt\right]. \tag{8}$$

The components of $\hat{\mathbf{b}} = (\hat{\mathbf{b}}_0, \hat{\mathbf{b}}_1)^T$ are replaced in equations (3) and (4) to estimate the parameters $\beta_0(t)$ and $\beta_1(t)$, respectively.

3 Relationship between rainfall curves in a region of Colombia

The methodology described is illustrated through its application to rainfall data recorded in weather stations from the Department of Valle del Cauca, Colombia in 2011. For each one of the 82 stations, we dispose of 73 pentadal rainfall data obtained from two sources of information (ground and satellite). These data were accumulated so that in the pentadal 73 (day 365 of the year) we have the accumulated rainfall of the year. Based on these data and using a B-splines basis (as smoothing method) we obtained for each station a curve (a functional datum) of accumulated rainfall (Figures 1 and 2). Assuming

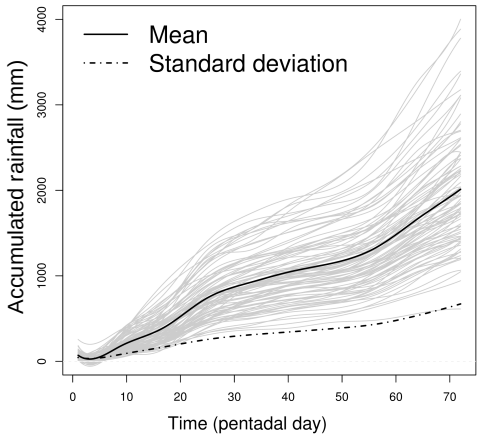


Figure 1: Ground accumulated rainfall curves (gray lines), ground accumulated rainfall mean curve (dark line) and ground accumulated rainfall standard deviation curve (dashed line). Curves were obtained by smoothing discrete accumulated rainfall data by using a B-splines basis. There are 82 curves (gray lines), each one corresponding to a meteorological station from the Department of Valle del Cauca, Colombia.

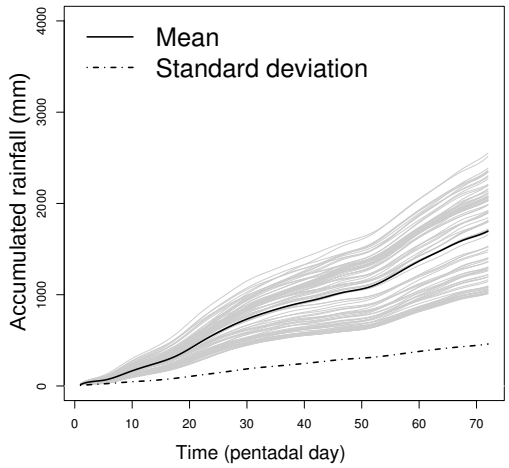


Figure 2: Satellite accumulated rainfall curves (gray lines), satellite accumulated rainfall mean curve (dark line) and satellite accumulated rainfall standard deviation curve (dashed line). Curves were obtained by smoothing discrete accumulated rainfall data by using a B-splines basis. There are 82 curves (gray lines), each one corresponding to a meteorological station from the Department of Valle del Cauca, Colombia.

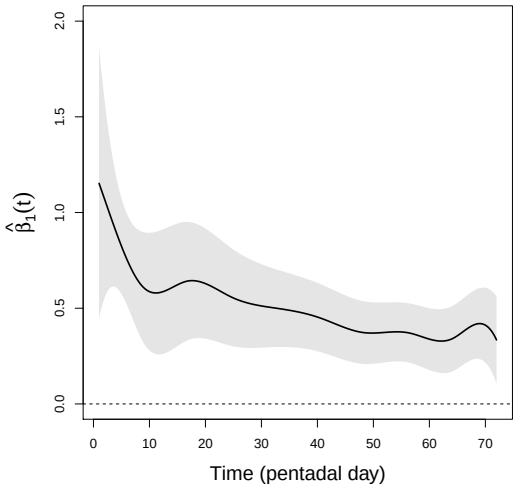


Figure 3: Estimated functional slope $\hat{\beta}_1(t)$ of the functional concurrent model (with spatially correlated errors) between ground and satellite accumulated rainfall curves (data from meteorological stations in Valle del Cauca, Colombia). In gray a 95% pointwise confidence band for $\beta_1(t)$.

independence, a functional regression model between ground and satellite curves was initially fitted. Then the functional residuals were used to explore the spatial correlation. The trace-variogram function was estimated (a exponential semivariance model was fitted). Once identified the spatial covariance structure of the errors, the covariance matrix Ω was estimated and used in equation (7) for obtaining the estimation by FLSE of $\mathbf{b}$. The components of this vector were posteriorly used in Equations (3) and (4) for estimating the parameters of interest. 95% point-wise confidence limits for the parameters are calculated following [4] by using the residuals of the final model (where we can assume independence between the errors). The estimated functional regression coefficient $\hat{\beta}_1(t)$ (Figure 3) shows a significant (the 95% confidence band do not include the zero in any period) positive contribution in the explanation of the ground rainfall data throughout the year. This relationship is much stronger at the start of the year (where there is less variability in both sets of curves). This estimation allows to conclude that the satellite information is highly related to the ground data but also indicates that a filter is required (for example a model such here estimated) if we want to use it in order of having a wide coverage of the precipitation in this region.

References

[1] N. Cressie, *Statistics for Spatial Data*. John Wiley and Sons, New York, 1993.

[2] R. Giraldo, *Geostatistics for functional data*, PhD Thesis, Universitat Politècnica de Catalunya, 2009

[3] R. Giraldo, P. Delicado, and J. Mateu, *Continuous time-varying kriging for spatial prediction of functional data: An environmental application*, Journal of Agricultural, Biological, and Environmental Statistics. 48(1) (2010), pp. 66–82.

[4] Ramsay, J. and Silverman, B. Functional Data Analysis, 2nd edition. 2005. New York: Springer.

[5] M. Schwaller and K. Morris, *A ground validation network for the global precipitation measurement mission*. Journal of Atmospheric and Oceanic Technology, 28 (2011), pp. 301–319.

Spatial and temporal drought variability in Tunisia

H. Feki[1*] and J. Carreau[2]

Ecole Supérieure des ingénieurs de Medjez el beb, Laboratoire GreenTeam, Univ. de Carthage, Tunisie; haifa.fki@gmail.com
[2] *HydroSciences Montpellier, Univ. de Montpellier, CNRS, IRD,France; julie.carreau@ird.fr*
[*] *Corresponding author*

Abstract. *Tunisia is suffering from intense and persistent drought episodes characterized by significant rainfall deficit. The country's historical memory confirms the abundance of drought sequences and their aggressiveness, particularly in arid zones. This study presents the interest of certain statistical and geostatistical methods for the spatial and temporal variability of the drought at different time steps. This drought would be characterized and quantified based on the triptych: "intensity, duration and geographical extent". The data used concern 67 raingauges covering the Tunisian territory and spreading over the period (1900-2015) and over which SPI indices are calculated. The PCA showed three similar areas in terms of drought, the average SPI for each region is then calculated. Spatial variability is analyzed based on the semi-variogram and the trend of the series is analyzed by the Mann Kendall test.*

Keywords. *Drought; Tunisia; Trend analysis; Geostatistics.*

1 Introduction

Tunisia has for years been suffering from intense and persistent drought episodes characterized by significant low rainfall amounts. Several studies and efforts have been made to research and analyze the causes and consequences of this variability in drought [1] and several drought indices have been used over the years, and all over the world, ranging from the simplest index of normal precipitation or precipitation percentiles, to the more complex, including the Palmer Drought Index. The specific objectives of the present study is: (1) to assess the local characterization of this phenomenon, by adopting the SPI like drought index (2) to detect trend in SPI time series, (3) to delimit homogenous regions by PCA and (4) to detect the spatial variability of drought using geostatistical methods.

2 Materials and methods

This study concerns all the Tunisian country. Tunisia is a transition climate zone located between the arid climate of the Sahara in the South and the Mediterranean humid climate to the North. Its climate is mostly semi-arid with hot dry summers and cold wet winters. In addition, high precipitation variability and topography repartition yield different natural climatic regions: humid in the North, semi-arid in the

Center and arid to the South. Collected data concern 67 rainfall stations at monthly time scale.

The different analysis tools that were used in this study are:

- The calculation of the standardized precipitation index SPI for 67 rainfall stations. The Standardized Precipitation Index (SPI) is a widely used index to characterize meteorological drought on a range of timescales ;
- Principal Component Analysis: to look for areas of similarity in terms of drought ;
- Analysis of trends and breaks in the SPI time series ;
- Spatial analysis using geostatistical based methods.

3 Results and discussion

T he aim of the PCA application is to detect homogenous regions in terms of drought. The study area was delineated into homogeneous sub-regions or zones of drought resemblance (ZRS) by grouping stations that were statistically correlated in terms of drought and reflecting different climatic and physiographic characteristics. The boundaries of the ZRS are delineated taking into account geographic features and the spatial distribution of precipitation. At this level, each ZRS is studied by calculating the average SPI index. For the 3 zones ZRC as shown in Fig. 1, the years of strong droughts are not identical until the first half of the 20th century, they do not generally coincide towards the end. In region C1, the SPI values indicate dry periods during 223 months for a percentage of 16% of the total time analyzed (= 115 years). There are 45 short-term dry events with an average duration of about 8 months. The most notable dry periods begin in November 1913 with a duration of 15 months, with a maximum value of (SPI-3 = -1.16) which indicates a moderate drought during this period. The longest and most deficient episode is between the months of March 1945 and June 1947 which reveals a deficit of cumulative rainfall for 28 consecutive months whose drought reached its extreme phase in April 1945 indicates a trough (SPI -3 = -2.61).

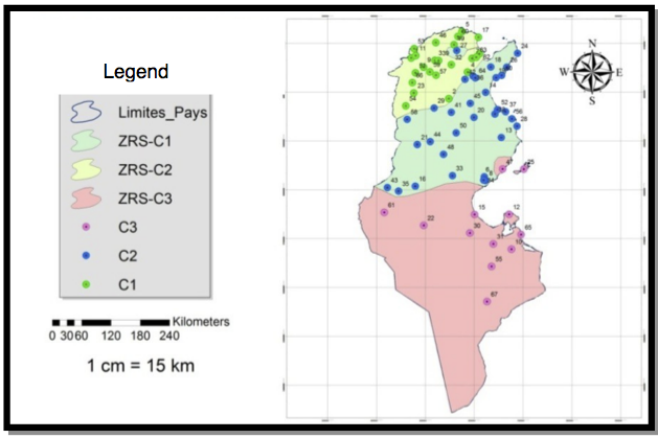


Figure 1: Homogenous regions delimitation PCA based

Exponential and pure nugget variogram models, see Fig. 2, are the most present ones for all directions. Pure nugget models occur generally in the middle of the 20th century indicating that the spatial structure of the phenomena is completely random.

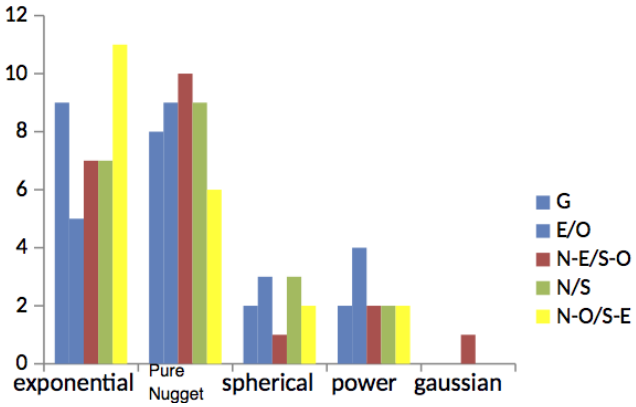


Figure 2: Variogram model occurence

4 Conclusion

The objective of this study was to demonstrate the potential of the SPI index to study the drought phenomenon via a number of analysis and mapping tools. Thus, our results show that drought is a phenomenon with great variability and which is greatly aggravated by the climate changes felt at the level of the Mediterranean. Using this drought index, we have achieved the ability to track spatio-temporally and analyze the evolution of drought.

References

[1] Ellouze, M. (2010). Caractérisation spatio-temporelle de la sécheresse en Tunisie centrale et méridionale et développement des hyétogrammes synthétiques, *mémoire de thèse de doctorat*, Université de Sfax.

Soil Organic Carbon modelling using jointly different sources

M. Zaouche[1,*], L. Bel[1] J. Tressou[1] and E. Vaudour[2]

[1] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay; mounia.zaouche@agroparistech.fr, liliane.bel@agroparistech.fr, jessica.tressou@agroparistech.fr*
[2] *UMR ECOSYS AgroParisTech, INRA, Université Paris-Saclay; emmanuelle.vaudour@agroparistech.fr*
[*] *Corresponding author*

Abstract. *Organic carbon is a good indicator of soil fertility and enables to mitigate gas emissions. Having at our disposal a precise mapping of its content is therefore essential. In this study we aim at spatially estimate the soil carbon content (SOC) in the Versailles plain and the Alluets plateau, a 221 km^2 agricultural area. The novel Bayesian inference approach called Integrated Nested Laplace Approximation with Stochastic Partial Differential Equation (INLA-SPDE) allows us to ensure consistency between various available sources of information (soil samples and optical satellite image) and to produce in a short time a posteriori estimations of the parameters and the SOC field, considered as a latent field. Two models were evaluated and compared using the elevation covariate stemming from a Digital Elevation Model (DEM), including or not the data from the satellite image. Adding the image improves the prediction quality in terms of RMSE (Root Mean Square Error RMSE) since the RMSE goes from* 4.48 $g.Kg^{-1}$ *to* 3.81 $g.Kg^{-1}$ *using a validation set of 75 locations. Overall the carbon prediction map from the joint model represents more realistically the spatial structure and variability of the carbon field.*

Keywords. *Spatial statistics; INLA-SPDE; SOC; Joint modelling.*

---

1 Introduction

Precise mapping for organic carbon is essential as it is a good indicator for soil fertility and gas emissions mitigation. We aim at predicting topsoil organic carbon (SOC) of bare cultivated soils over the Versailles plain and the Alluets plateau, a 221 km^2 agricultural area with both contrasted soils and SOC contents, located in the western region of Paris. In a previous study ([4]), we produced SOC predictions maps using geostatistical techniques based on soil samples and an elevation covariate stemming from a 25-m of resolution Digital Elevation Model (DEM). Nevertheless, there are limitations using only field samples: field measurements are expensive, limited in number and not sufficient enough to give a precise idea of the variations on a local scale. In another study over this same region, reflectance spectroscopy has been used as cheaper alternative to measure SOC content ([3]). The approach here is to propose a model ensuring consistency between two available sources of information with different nature, namely field samples and a SPOT4 image of a 20-m resolution in a Bayesian framework. Given the large size of the data sets, we resort to the The Bayesian inference Integrated Nested Laplace Approximation (INLA) combined with Stochastic Partial Differential Equation (SPDE) which is mostly used in spatial modelling in case of huge data sets ([1],[2]). It allowed us to produce accurate estimations of the parameters and

the SOC field, considered as a latent field. In this study, we have compared predictive performances of two models both using the elevation covariate from the DEM: one integrating only field observations, the other one including field observations and reflectances measurements jointly. Based on 253 field samples collected from 2010 to 2013, this comparison has been made in two ways: through a resampling bootstrap procedure (40 repetitions, 70 locations for validation) first, then using a combination of two extra soil samples collected in 2016 and 2017 as validation sets.

2 Data sets

From 2010 to 2013, soil samples were collected at 253 locations sites, in 2016 at 26 sites, and in 2017 at 49 sites in the topsoil layer (Figure [1], left). We have used the locations from 2010 to 2013 as the calibration set and combinations of 2016 and 2017 as validation sets. A SPOT4 image with 20m resolution was acquired over the study zone on 17 April 2013. This multispectral image consists of 4 spectral bands in the green, red, and infrared domains. We kept only its first one b1 ($500-590$ *nm*), as the remaining bands carried the same type of information. Removing the vegetated soils led us to exploit $105,941$ pixel measurements. Focusing on the relation between the SOC contents and associated reflectances, we observe it is shaped more like an exponential link rather than a linear link.

3 Model definition and inference approach

Model based on field samples

This model (denoted $Elev^1$) is set up as follows:

$$y_i^S \sim Normal(\eta_i^S, \varepsilon_i^S)$$

with $\eta_i^S = \beta_0 + \beta_1 E_i + \xi_i$ the linear predictor. y_i^S is the SOC content sampled at site i, ε^S is a Gaussian white noise with variance $\sigma^2_{\varepsilon_s}$, β_0, β_1 are the fixed effects parameters, E_i is the elevation value at location i, ξ is a Gaussian field with a Matérn covariance function with scaling parameter κ_ξ, smoothness parameter ν and variance σ^2_ξ.

Joint model: field samples and image reflectance measurements

The spatial range previously estimated (500m) in [4] is much greater than the data grid resolution (20m). We henceforth consider the surface reflectances of individual pixels as locations. Given the relation between the reflectance data and the SOC content measurements, we assume that reflectance is an exponential distributed variable. Both sources share the same spatial term ξ and an additive spatial term γ is used in the reflectance modelling. The joint model (denoted $SPOTb1-Elev^{12}$) is defined as follows :

$$y_i^S \sim Normal(\beta_0 + \beta_1 E_i + \xi_i, \varepsilon_i^S)$$
$$y_j^R \sim Exponential(\lambda_j)$$

with $\eta_i^S = \beta_0 + \beta_1 E_i + \xi_i$ and $\eta_j^R = log(1/\lambda_j) = \beta_2 + \beta_3 E_j + \theta_0 \xi_j + \gamma_j$ are the linear predictions. The linear predictor here is defined as $\eta_{joint} = \begin{pmatrix} \eta^S \\ \eta^R \end{pmatrix}$.

y_j^R is the reflectance value at site j, β_3 and β_4 are fixed effects parameters and κ_γ, ν, σ_γ^2 are the Matérn parameters of γ.

Inference with INLA-SPDE

The INLA algorithm ([2]) allows us to perform the inference of hierarchical models with Gaussian latent processes. These models consist of three layers: hyperparameters θ, latent vector x, and a likelihood model. If $\xi = \{\xi(s), s \in D \subseteq \mathbb{R}^2\}$ is a random field with a Matérn covariance function characterized by scaling and smoothing parameters κ and ν, the objective of the SPDE approach is to find a Gaussian Markov Random Field (GMRF) $\tilde{\xi}$ with a local neighborhood and a sparse precision matrix, which approximates the field ξ. An approximate solution of the SPDE is obtained by dividing the domain into a set of non-intersecting triangles (Figure[1]).

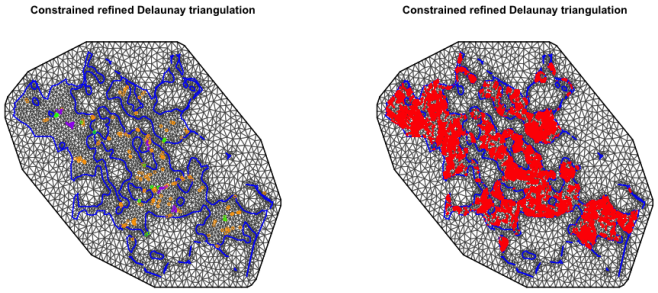


Figure 1: Triangulation of the study area. Left: samples locations collected, from 2010 to 2013 (orange), in 2016 (green), in 2017 (purple). Right: pixel centers locations (red) coinciding with bare soils.

Both models were implemented using the R-package R-INLA, keeping the default value of the parameter ν (e.g equal to 1) for both. In $Elev^1$, the latent vector is $x = (\eta^S, \tilde{\xi}, \beta_0, \beta_1)$ and the hyperparameter is $\theta = (\kappa_\xi, \sigma_\xi^2, \sigma_{e_s}^2)$. In $SPOTb1 - Elev^{12}$, the latent vector is $x = (\eta_{joint}, \tilde{\xi}, \tilde{\xi}^c, \tilde{\gamma}, \beta_0, \beta_1, \beta_2, \beta_3)$ with $\tilde{\gamma}$ the GMRF approximating γ and $\tilde{\xi}^c = \theta_0 \xi$, and the hyperparameter is $\theta = (\theta_0, \kappa_\xi, \sigma_\xi^2, \sigma_{e_s}^2, \kappa_\gamma, \sigma_\gamma^2)$.

4 Results and perspectives

The model $SPOTb1 - Elev^{12}$ has slightly improved the prediction quality in the bootstrap procedure (Table 1).

Model	Mean	Median	Min	Max
$Elev^1$	3.23	3.17	2.78	4.00
$SPOTb1 - Elev^{12}$	3.22	3.15	2.53	4.98

Table 1: Bootstrap RMSE statistics in $g.kg^{-1}$.

The difference is favourably wider when using combinations of 2016 and 2017 validation sets, whatever the data set (Table 2).

	RMSE ($g.Kg^{-1}$)		
calibration set (size)	2010-2013 (253)	2010-2013 (253)	2010-2016 (279)
validation set (size)	2016-2017 (75)	2016 (26)	2017 (49)
$SPOTb1 - Elev^{12}$	3.81	3.76	3.96
$Elev^{1}$	4.49	3.98	4.62

Table 2: RMSE results in $g.Kg^{-1}$ using the 2016 and 2017 validation sets.

In addition, looking at the mean square error at each site revealed that $SPOTb1 - Elev^{12}$ could carry information for isolated locations and more realistically depict the SOC content variability. The comparison of the SOC content prediction maps we have produced for both models confirmed this analysis (Figure [2]).

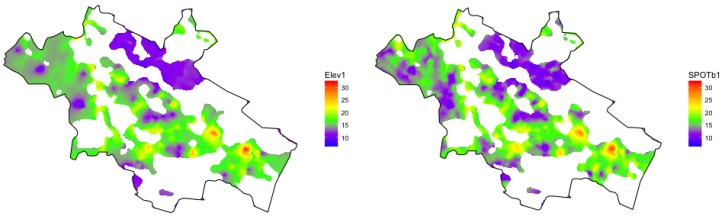


Figure 2: SOC prediction map. Left: $Elev^{1}$. Right: $SPOTb1 - Elev^{12}$.

Further work will be to determine the threshold number of field samples from which $SPOTb1 - Elev^{12}$ is to be preferred to $Elev^{1}$. Moreover, another perspective would be to build models that could integrate the information carried by all the layers stemming from multispectral images.

Acknowledgments. This study was carried out in the framework of the TOSCA PLEIADES-CO project supported by the French Space Agency (CNES).

References

[1] Blangiardo, M., Cameletti, M., Baio, G., Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology* **4**, 33–49.

[2] Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*. **71(2)**, 319–392.

[3] Vaudour, E., Bel, L., Gilliot, Coquet, Y. (2013). Potential of SPOT multispectral satellite images for mapping topsoil organic content over a peri-urban croplands. *SSAJ* **77**, 2122–2139.

[4] Zaouche, M., Bel, L., Vaudour, E. (2017). Geostatistical mapping of topsoil organic carbon and uncertainty assessment in Western Paris croplands (France). *Geoderma Regional* **10**, 126–137.

Spatial-temporal pattern analysis and prediction of air quality using Discrete Fourier Transform

L.Ippoliti and E.Nissi*

Department of Economics -University G.d'Annunzio Chieti Pescara Italy;
*luigi.ippoliti@unich.it; eugenia.nissi@unich.it, *Corresponding author*

Abstract. *Spatio-temporal modelling has received more and more attention from academic researchers for its promising applicability to complex data containing both spatial and temporal patterns. In this work, we discuss the modelling of Particulate Matter (PM10) data by using a frequency domain approach. We show that our model has several computational advantages, and that it is able to provide good predictions and can be used for descriptive purposes.*

Keywords. *Discrete Fourier Transform; Spatio-temporal Processes; Prediction; Environmental Data*

1 Introduction

Spatio-temporal statistics has been drawing more and more attention from academic researchers and industrial practitioners for its promising applicability to complex data containing both spatial and temporal characteristics.
Different types of models have been proposed for analysing such data. A commonly used approach is to work with models having a directly specified correlation structure, such as those used in geostatistics – see, for example, Gneiting (2002); Stein (2005) and Rodrigues and Diggle (2010). The spatio-temporal models developed in this direction view time as continuous rather than discrete and more emphasis is put on spatial prediction but less on forecasting future values. Another possibility is to follow a dynamic modelling approach by means of state space models. Example on this line are given, for example, by Mardia et al. (1998) and Cressie and Wikle (2011).
Differently from above, in this work, we discuss a spatio-temporal model which is developed by using a frequency domain approach. By discussing an application on air pollution, we show that our model has several computational advantages, and that it is able to provide good predictions and can be used for spatially descriptive purposes.

1.1 The model

Let $Y_t(s)$, where $\{\mathbf{s} \in \mathbb{R}^d,\ t \in \mathbb{Z}\}$, denote a spatio-temporal random process. We assume that the random process is spatially and temporally second order stationary, i.e. $E[Y_t(\mathbf{s})] = 0$, $Var[Y_t(\mathbf{s})]$=σ^2_Y and $Cov[Y_t(\mathbf{s})Y(t+u)](\mathbf{s}+\mathbf{h})] = c(\mathbf{h},u)$, $\mathbf{h} \in \mathbb{R}^d,\ u \in \mathbb{Z}$. With no loss of generality assume that the process

has zero mean. The spectral density function of $Y_t(\mathbf{s})$ is defined as

$$f(\underline{\lambda},\omega) = \frac{1}{(2\pi)^{d+1}} \sum_u \int_{-\infty}^{+\infty} e^{-i(\mathbf{h}'\underline{\lambda}+u\omega)} c(\mathbf{h},u)\, d\mathbf{h} \tag{1}$$

and its inverse relationship is

$$c(\mathbf{h},u) = \int_{-\infty}^{+\infty} \int_{-\pi}^{+\pi} e^{i(\mathbf{h}'\underline{\lambda}+u\omega)} f(\underline{\lambda},\omega)\, d\underline{\lambda}\, d\omega. \tag{2}$$

Let $\{Y_t(\mathbf{s}_i)|i=1,2,\ldots,m;t=1,2,\ldots,n\}$ be a sample from the zero mean temporal stationary process $\{Y_t(\mathbf{s}_i)\}$ at the location $\mathbf{s}_i$, and define the Discrete Fourier transform (DFT)

$$J_{s_i}(\omega_k) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} Y_t(\mathbf{s}_i)\, e^{-it\omega_k} \tag{3}$$

where $\omega_k = \frac{2\pi k}{n}$ $k = 0, \pm ..., \pm[\frac{n}{2}]$. In practice one uses Fast Fourier Transform algorithm to compute the DFT of the time series observed at site $\mathbf{s}_i$. From the above, by inversion, we get

$$Y_t(s_i) = \sqrt{\frac{n}{2\pi}} \int_{-\pi}^{\pi} J_{s_i}(\omega) e^{it\omega}\, d\omega. \tag{4}$$

By following Subba Rao and Terdik (2017), it can be shown that a valid spectral density function for the spatio-temporal process is given by

$$f(\underline{\lambda},\omega) = \frac{\sigma_e^2}{2\pi^3(\lambda_1^2+\lambda_2^2+|c(\omega)|^2)^{2\nu}} \tag{5}$$

where $|c(\omega)|^2$ is specified below. If the stationary spatio-temporal process is isotropic, then the covariance function between the discrete Fourier Transforms $J_s(\omega)$ and $J_{s+h}(\omega)$ is obtained by the inverse Fourier transform of equation (5) which gives

$$g_{\|\mathbf{h}\|}(\omega) = Cov(J_s(\omega), J_{s+h}(\omega)) = \frac{\sigma_e^2}{(2\pi)^2} \Big(\frac{\|\mathbf{h}\|}{2|c(\omega)|}\Big)^{2\nu-1} \frac{K_{2\nu-1}(|c(\omega)|\ \|\mathbf{h}\|)}{\Gamma(2\nu)} \tag{6}$$

where $K_\alpha(.)$ is the modified Bessel function of the second kind of order α. An interesting feature of this covariance function is that the argument of the Bessel function derived above is not only a function of the spatial distance, but also a function of the frequency dependent scaling function, which is related to the second order temporal spectral density function. For our process, under the conditions stated above, it can be shown that this temporal spectral density function is obtained by considering the limiting behavior of $g_{\|\mathbf{h}\|}(\omega)$ as$\|\mathbf{h}\| \to 0$, which results

$$g_0(\omega) = \frac{\sigma_e^2}{2(2\pi)^2(|c(\omega)|^2)^{2\nu-1}(2\nu-1)}. \tag{7}$$

If one considers the special case of $\nu = 1$, we clearly note that $|c(\omega)|^2$ is proportional to the inverse of the spectral density function $g_0(\omega)$ which, in turn, thus affects the behaviour of the covariance function $g_{\|\mathbf{h}\|}(\omega)$.

Clearly, given $g_{\|\mathbf{h}\|}(\omega)$, it is possible to rely on the kriging framework to predict the Fourier coefficients, $J_{S_0}(\omega)$, at a new site s_0 and, accordingly, obtain the predicted data by means of the inverse transform

$$\widehat{Y}_t(s_0) = \sqrt{\frac{n}{2\pi}} \int_{-\pi}^{\pi} \widehat{J}_{S_0}(\omega) e^{it\omega} d\omega. \tag{8}$$

1.2 Real Data Application

This section provides some details of a preliminary analysis carried out on daily PM_{10} time series collected at 70 monitoring stations in Lombardia (Italy) during the year 2011. The monitoring network is shown in Figure 1 while the data are shown in Figure 2.

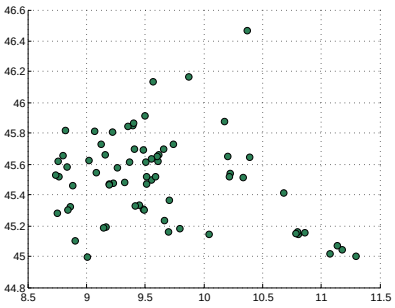


Figure 1: Pm_{10} Monitoring stations operating in Lombardia region in 2011

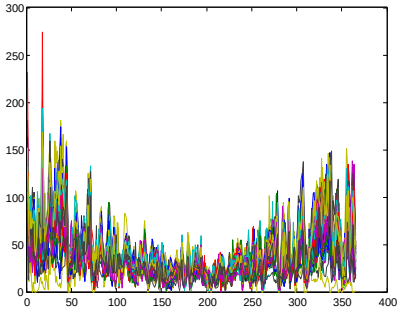


Figure 2: Daily time series concentrations of PM_{10}

We have assumed $\nu = 1$ and predicted the detrended data by using a cross-validation procedure where each time series is excluded in turn ad used only for prediction purposes. An exploratory analysis of the data suggests that an AR(1) model may be adequate to explain the temporal dependence and specify the structure of $|c(\omega)|^2$. Results suggest that the model is able to predict the process satisfactorily. As expected, the best results are obtained especially for sites at the center of the configuration for which the number of neighbours is larger. As an example, Figure 3 shows typical prediction results for sites with coordinates ranging between $45.8 - 45.4$ and $9.5 - 9.8$ for latitude and longitude, respectively.

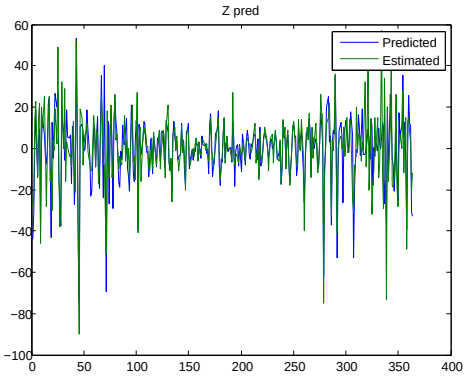


Figure 3: Plot of the observed (detrended data) and predicted values for a monitoring site

References

[1] Cressie, N.,Wikle C.K.,(2011). Statistics for Spatial Data. *John Wiley* .

[2] Gneiting, T.,(2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association,* **97**, 590–600.

[3] Mardia, K.V.,Goodall C., Redfern E.,Alonso F.J.,(1998). The Kriged Kalman Filter. *Test,* **7**, 217–285.

[4] Paci L., T. , Finazzi F.,(2018). Dynamic model-based clustering for spatio-temporal data. *Stat Comput,* **28**, 359–374.

[5] Rodriguez, A., Diggle P.J.(2010). A Class of Convolution Based Models for Spatio-Temporal Processes with Non Separable Covaraince Structure. *Scandinavian Journal of Statistics,* **37**, 553–567.

[6] Stein., M L, ,(2005). Spacetime covariance functions. *ournal of the American Statistical Association* **100**, 310–321.

[7] Subba Rao, T. ,Das,S.,Boshnakov.,G.A, (2014). A frequency domain approach for the estimation of parameters of spatio-temporal random processes. *Journal of Time Series Analysis,* **35**, 357–377.

[8] Subba Rao, T. Terdik, Gy,(2017).A New Covariance Function and Spatio-Temporal Prediction (Kriging) for A Stationary Spatio-Temporal Random Process , *Journal of Time Series Analysis,* **38**, 936–959.

Spatio-Temporal Modelling of Criminal data in Portugal

Conceição Ribeiro[1,3,4,*] and Paula Pereira[2,3]

[1] *Instituto Superior de Engenharia, Universidade do Algarve; cribeiro@ualg.pt*
[2] *Escola Superior de Tecnologia de Setúbal, Instituto Politécnico de Setúbal; paula.pereira@estsetubal.ips.pt*
[3] *Centro de Estatística e Aplicações, Universidade de Lisboa*
[4] *CEPAC, Universidade do Algarve*
[*] *Corresponding author*

Abstract. *The study of the evolution of crime, whether in a temporal level or in a space level, presents a great importance in the definition of measures to improve the welfare of the population. Usually, to analyze the evolution of crime in a given region, it resorts to compare rates of several years. This work aims to extend this analysis and use spatial and temporal models that allow to characterize the trend of crime in the spatial level and in the temporal level. These models are applied to data crime observed in the municipalities of mainland Portugal, from 2011 to 2016.*

Keywords. *Crime trends; Spatio-temporal; Bayesian modelling; Small area.*

---

1 Introduction

In 1996 the World Health Organization (WHO) adopted Resolution WHA49.25 – Preventing violence: a public health priority, and declared "violence a leading worldwide public health problem". With this resolution it was intended to call attention to the severity of the consequences of violence for individuals, families, communities and countries, and highlight the damaging effects of violence on public health [8]. Moreover the costs of violence to a nation's economy need to be taken into consideration which includes the direct costs of medical care and criminal justice as well as indirect costs. The public health approach is science-based and multidisciplinary and complements criminal justice and human rights responses to violence. Nevertheless, action requires measuring violence which presents many challenges but certainly allows the vital basis knowledge for policy-making. Reliable information on violence is fundamental for planning and monitoring purposes and there are several sources of information [8]. In this perspective the study of the evolution of crime, whether in a temporal level or in a space level, presents a great importance in the definition of measures to improve the welfare of the population thus contributing to the public health approach to violence prevention. To analyze the evolution of crime in a given region usually it resorts to compare rates of several years. This work aims to extend this analysis and use spatio-temporal models that allow characterizing the trend of crime in the spatial level and in the temporal level. In other words, it intends to understand if over the years and across regions there have been changes in crime patterns. Also it aims to study the influence of population characteristics through some related covariates.

2 Study Region and Data

The study region consists on the 278 municipalities of continental Portugal. Portugal has an area of 89.015 km^2 and a population of approximately 10.000.000 inhabitants. The majority of the population lives in urban areas and near the litoral coast. The data consists of annual number of crimes recorded by the police authorities by geographic localization and was obtained from the official statistics published by the National Statistical Institute (INE) of Portugal. The crimes analyzed where divided in two categories:

- Crimes against persons: crimes against life, crimes against physical integrity, crimes against personal freedom, crimes against freedom and sexual self-determination, crimes Against Honor and Crimes against privacy.
- Crimes against patrimony: crimes against property, crimes against property in general and crimes against property rights.

3 Methodology

Several spatio-temporal Bayesian models of the field of epidemiology were considered and the model of Bernardinelli et al. [1] was used because it considers terms that allow to identify the mean trend and area-specific trends as well as to evaluate whether there is spatio-temporal interaction effects. Furthermore, the model stabilizes area-specific risk estimates when data are scarce in small areas. These advantages allow crime analysis to be conducted at a large map scale with small areal unit and provide insight into the local distribution and patterns of crime trend.

Thus hierarchical Bayesian models were used and to implement these models INLA methodology (*Integrated Nested Laplace Approximations*) was used through the package of R, R-INLA, [6]. The models were applied to crime data observed in the 278 municipalities of the continental Portugal, from 2011 to 2016. In this way

$$y_{it}|\theta_{it} \sim Poisson(E_{it}\theta_{it}), \qquad i = 1,...,278, \quad t = 1,...,6$$

where y_{it} is the annual crime number in municipality i and in year t, θ_{it} the relative risk in municipality i and in year t, E_{it} the expected number of crimes in municipality i and in year t, and n_{it} is the number of inhabitants in municipality i and in year t.

The relative risk is defined as

$$\log(\theta_{it}) = \alpha + \beta X_{it} + s_i + (\gamma + \delta_i) \times t, \qquad i = 1,...,278, \quad t = 1,...,6$$

where α and βX_{it} are fixed effects (α is the intercept, X_{it} are the covariates and β are the covariates coefficients); s_i are the spatial effects; $(\gamma + \delta_i) \times t$ are the temporal effects (γ is the global linear trend average and δ_i are the spatio-temporal interaction random effects).

Since this is a preliminary study of crime trends, the covariates included in the model are not intended to establish a possible association between risk factors and crime trends. In this preliminary study only three covariates were included regarding economic aspects: purchasing power (x_{it1}), beneficiaries of social integration income of social security per 1000 inhabitants in active age (x_{it2}) and dynamism

relative factor of purchasing power (x_{it3}). In a future work we intend to include covariates related to social aspects in order to establish possible associations between risk factors and crime trends.

The specification of priors for the parameters were: $\alpha, \beta_1, \beta_2, \beta_3 \sim Normal(0, 1000)$, $s_i \sim CAR(\sigma_s^2)$, $\gamma \sim Normal(0, 1000)$, $\delta_i \sim CAR(\sigma_\delta^2)$, for σ_s^2 and σ_δ^2 penalised complexity priors were used [7]. Model fit and model selection was evaluated using Deviance Information Criterion (DIC) and Watanabe-Akaike information criterion (WAIC).

4 Results

The results in Table 1 and Figure 2 show a decreasing temporal trend of crime from 2011 to 2016 in all types of crime studied in this work ($\gamma < 0$), indicating that crimes against persons and crimes against patrimony have had a decreasing trend from 2011 to 2016. In the selected model, the variation due to spatial correlation (σ_s^2) is greater than the variation due to spatio-temporal interaction (σ_δ^2). This implies that the largest influence on the mean trend is spatial structured although there is an area specific differential trend across the study region. Moreover, the covariate "dynamism relative factor of purchasing power" was not included in the selected model for crimes against persons, since it was not significant.

	Mean	St.Dev.	2.5%CI	97.5%CI
Fixed effects:				
α	−0.3526	0.0524	−0.4554	−0.2497
β_1	0.3404	0.0591	0.2241	0.456
β_2	2.8308	0.4255	1.9953	3.6655
γ	−0.0186	0.0020	−0.0225	−0.014
Random effects variance:				
σ_s^2	0.1373	0.0137	0.1127	0.1661
σ_δ^2	0.0021	0.0003	0.0016	0.0028

	Mean	St.Dev.	2.5%CI	97.5%CI
Fixed effects:				
α	−0.3995	0.0413	−0.4806	−0.3184
β_1	0.3738	0.0459	0.2836	0.4639
β_2	0.9964	0.3362	0.3359	1.6556
β_3	0.0306	0.0079	0.0152	0.0461
γ	−0.0681	0.0016	−0.0712	−0.0650
Random effects variance:				
σ_s^2	0.2376	0.0219	0.197	0.2835
σ_δ^2	0.0053	0.0006	0.0043	0.0066

Table 1: *Posterior* summaries of the parameters: Crimes against persons (left) and Crimes against patrimony (right).

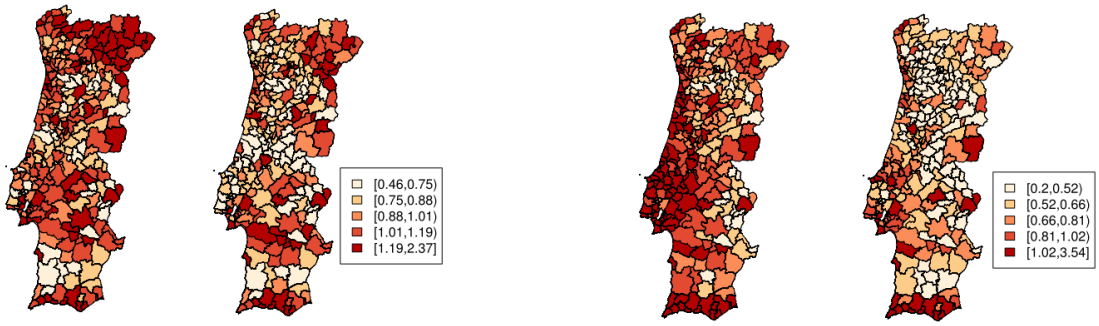


Figure 1: Left (the first two maps): *Posterior* relative risk estimates of the number of Crimes against persons: 2011 and 2016. Right (the last two maps): *Posterior* relative risk estimates of the number of Crimes against patrimony: 2011 and 2016.

Hot spots are defined as areas that have high probabilities of *Posterior* relative risk being greater than one. The results for both types of crimes and for year 2016 are showned in Figure 2. Thus regarding Crimes against persons, the northeast and the south of Portugal can be considered as a hot spot and, in the case of Crimes against patrimony, the south of Portugal can be consider as a hot spot.

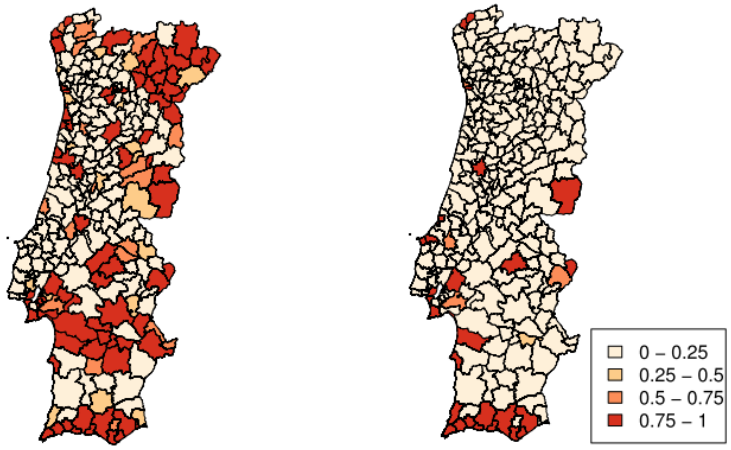


Figure 2: Probabilities of *Posterior* relative risk of 2016 being greater than one: Crimes against persons (left) and Crimes against patrimony (right).

Acknowledgments. This work is partially sponsored by national funds through FCT - Fundação para a Ciência e a Tecnologia under the project UID/MAT/00006/2013.

References

[1] Bernardinelli L., Clayton D., Pascutto C., Montomoli C., Ghislandi M., Songini M. (1995) Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine*, **14**, 2433–2443.

[2] Knorr-Held L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics In Medicine*, **19(17–18)**, 2555–2567.

[3] Law J., Quick M., Chan P. (2014). Bayesian Spatio-Temporal Modeling for Analysing Local Patterns of Crime Over Time at the Small-Area Level. *Journal of Quantitative Criminology*, **30**, 57–78.

[4] Law J., Quick M., Chan P. (2015). Analyzing Hotspots of Crime Using a Bayesian Spatiotemporal Modeling Approach: A Case Study of Violent Crime in the Greater Toronto Area. *Geographical Analysis*, **47(1)**, 1–10.

[5] Ribeiro C., (2013). *Modelação de Dados Espaço Temporais em Segurança Rodoviária*. Phd tesis, Universidade de Lisboa, Portugal.

[6] Rue H., Martino S., Chopin N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, **71(2)**, 319–392.

[7] Simpson D.P., Rue H., Martins T.G., Riebler A., Sørbye S.H. (2014). *Penalising model component complexity: A principled, practical approach to constructing priors*. ArXiv e-prints.

[8] WHO (2002), *World report on violence and health*.
`http://www.who.int/violence_injury_prevention/violence/world_report/en/`