NONPARAMETRIC GEOSTATISTICAL RISK MAPPING

R. Fernández Casal[1], M. Francisco-Fernández[1], A. Quintela-del-Río[1], S. Castillo-Páez[2]

[1] University of A Coruña (Spain); ruben.fcasal@udc.es, mariofr@udc.es, aquintela@udc.es

[2] University of Vigo (Spain); sacastillo@uvigo.es

Abstract

In this work, a fully nonparametric geostatistical approach to estimate threshold-exceeding probabilities is proposed. We suggest to use the nonparametric local linear regression estimator, with a bandwidth selected by a method that takes the spatial dependence into account, to estimate the large-scale variability (spatial trend) of a geostatistical process. To estimate the small-scale variability, a bias-corrected nonparametric estimator of the variogram is proposed. Finally, a bootstrap algorithm is used to estimate the probabilities of exceeding a threshold value at unsampled locations. The behavior of this approach is also evaluated through simulation and with an application to a real data set.

Introduction

- Assuming that $\left\{Y(\mathbf{x}), \mathbf{x} \in D \subset \mathbb{R}^d\right\}$ is a spatial process that can be modeled as:

$$Y(\mathbf{x}) = m(\mathbf{x}) + \varepsilon(\mathbf{x}), \qquad (1)$$

  where $m(\cdot)$ is the trend function and the error term ε, is a second order stationary process with zero mean and covariogram $C(\mathbf{h}) = Cov(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{x}+\mathbf{h}))$.

- In this framework, given n observed values $\{Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n)\}$, the goal is, using a fully nonparametric geostatistical approach, to estimate the conditional probability:

$$P\left(Y(\mathbf{x}_0) \geq c | \mathbf{Y}\right)$$

  where c is a threshold (critical) value and $\mathbf{Y} = (Y(\mathbf{x}_1), ..., Y(\mathbf{x}_n))^t$.

- There are several geostatistical approaches which may be used to estimate the probabilities of exceeding a threshold value, such as indicator kriging, disjunctive kriging or Markov chain geostatistical modeling, among others (see, e.g. [7]). However, the results obtained when these procedures are applied in practice could be unsatisfactory, usually due to the misspecification of the assumed parametric model (apart from other potential issues).

- In this work, under the general spatial model (1), and without assuming any parametric model for the trend function and for the dependence structure of the process, a general nonparametric procedure for spatial risk assessment is proposed.

- This procedure is a modification of the semiparametric bootstrap described in [6], which tries to adequately reproduce the variability of the data.

Nonparametric geostatistical modeling

- The local linear trend estimator (e.g. [8]) is given by:

$$\hat{m}_{\mathbf{H}}(\mathbf{x}) = \mathbf{e}_1^t \left(\mathbf{X}_{\mathbf{x}}^t \mathbf{W}_{\mathbf{x}} \mathbf{X}_{\mathbf{x}}\right)^{-1} \mathbf{X}_{\mathbf{x}}^t \mathbf{W}_{\mathbf{x}} \mathbf{Y} \equiv s_{\mathbf{x}}^t \mathbf{Y},$$

  where e_1 is a vector with 1 in the first entry and all other entries 0, $\mathbf{X}_{\mathbf{x}}$ is a matrix with ith row equal to $(1, (\mathbf{x}_i - \mathbf{x})^t)$, $\mathbf{W}_{\mathbf{x}} = \mathtt{diag}\left\{K_{\mathbf{H}}(\mathbf{x}_1 - \mathbf{x}), ..., K_{\mathbf{H}}(\mathbf{x}_n - \mathbf{x})\right\}$, $K_{\mathbf{H}}(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$, K is a multivariate kernel and $\mathbf{H}$ is a $d \times d$ nonsingular symmetric matrix.

- The bandwidth matrix $\mathbf{H}$ controls the shape and size of the local neighborhood used to estimate $m(\mathbf{x})$. We recommend the use of the "bias corrected and estimated" generalized cross-validation (GCV) criterion proposed in [5] to select this bandwidth in practice.

- As in traditional geostatistical approaches, the usual dependence estimation method consists in removing the trend and estimating the variogram from the residuals:

$$\hat{\varepsilon} = \mathbf{Y} - \mathbf{SY}$$

- Nevertheless, it is well-known that the direct use of the residuals in the estimation of the variogram (or the covariogram) may produce a strong underestimation of the small-scale variability of the process (e.g. [2], Section 3.4.3). Simply note that:

$$Var(\hat{\varepsilon}) = \mathbf{\Sigma} + \mathbf{S\Sigma S}^t - \mathbf{\Sigma S}^t - \mathbf{S\Sigma} = \mathbf{\Sigma}_{\hat{\varepsilon}}$$

  where $\mathbf{\Sigma}$ is the covariance matrix of the errors.

- As this bias may also have a significant impact on the estimation of threshold-exceeding probabilities, a similar approach to that described in [3] will be used. Using an iterative algorithm, the squared differences of the residuals are conveniently corrected and used to compute a pilot local linear variogram estimate.

- The final variogram estimate is obtained by fitting a "nonparametric" isotropic Shapiro-Botha variogram model [9], to the bias-corrected nonparametric pilot estimate.

Bootstrap algorithm

The proposed algorithm is as follows:

1. Using the procedures described in previous section:
   (a) Obtain the optimal bandwidth matrix $\mathbf{H}$, the residuals $\hat{\varepsilon}_i = Y(\mathbf{x}_i) - \hat{m}_{\mathbf{H}}(\mathbf{x}_i)$, $i = 1, \ldots, n$, and the estimated covariance function $\hat{C}$ of the errors.
   (b) Using $\hat{C}$, compute the (estimated) covariance matrix of the errors $\hat{\mathbf{\Sigma}}$ and find the matrix $\mathbf{L}$, such that $\hat{\mathbf{\Sigma}} = \mathbf{LL}^t$, using Cholesky decomposition.
   (c) Compute the estimated covariance matrix of the residuals $\hat{\mathbf{\Sigma}}_{\hat{\varepsilon}}$, and the matrix $\mathbf{L}_{\hat{\varepsilon}}$, such that $\hat{\mathbf{\Sigma}}_{\hat{\varepsilon}} = \mathbf{L}_{\hat{\varepsilon}}\mathbf{L}_{\hat{\varepsilon}}^t$.
2. Generate a bootstrap sample as follows:
   (a) Compute the "independent" variables $\mathbf{e} = \mathbf{L}_{\hat{\varepsilon}}^{-1}\hat{\varepsilon}$.
   (b) These variables are centered and, from them, we obtain an independent bootstrap sample of size n, denoted by $\mathbf{e}^*$.
   (c) Next, the bootstrap errors $\hat{\varepsilon}^* = \left(\hat{\varepsilon}_1^*, \ldots, \hat{\varepsilon}_n^*\right)^t$ are $\hat{\varepsilon}^* = \mathbf{Le}^*$, and the bootstrap samples are $Y^*(\mathbf{x}_i) = \hat{m}_{\mathbf{H}}(\mathbf{x}_i) + \hat{\varepsilon}_i^*$, $i = 1, 2, \ldots, n$.
3. Compute the kriging prediction $\hat{Y}^*(\mathbf{x}_0)$ at each unsampled location $\mathbf{x}_0$ from the bootstrap sample $\{Y^*(\mathbf{x}_1), \ldots, Y^*(\mathbf{x}_n)\}$.
4. Repeat steps 2 and 3 a large number of times B (e.g. $B = 1,000$).

Finally, a map with the frequencies (across bootstrap replicates) of how often a location is included in the at-risk area is computed.

Simulation results

- $N = 1,000$ samples of different sizes were generated following model (1) on a regular grid in the unit square, with mean function $m(x_1, x_2) = \sin(2\pi x_1) + 4(x_2 - 0.5)^2$ and random errors ε_i normally distributed with zero mean and isotropic exponential covariogram:

$$\gamma_\theta(\mathbf{u}) = c_0 + c_1\left(1 - \exp\left(-3\frac{\|\mathbf{u}\|}{r}\right)\right),$$

  (for $\mathbf{u} \neq \mathbf{0}$), where c_0 is the nugget effect, c_1 is the partial sill and r is the practical range.

- The values considered in the simulations were:
  - Sample sizes of $n = 10 \times 10$, 17×17 and 20×20.
  - Sill (variance) values of $\sigma^2 = 0.16$ and 1.
  - Practical ranges of $r = 0.15$ and 0.5.
  - Nugget values of 0%, 25% and 50% of σ^2 ($c_1 = \sigma^2 - c_0$).

- Using the proposed procedure, estimated probabilities of obtaining a value of the response variable larger than a threshold c, with $c = 2.0, 2.5, 3.0$ and 3.5, were computed in a 50×50 regular grid. For instance, Figure 1 shows the maps with theoretical and estimated probabilities for $c = 2.5$, $n = 20 \times 20$, $\sigma^2 = 0.16$, $r = 0.5$ and $c_0 = 0.04$.

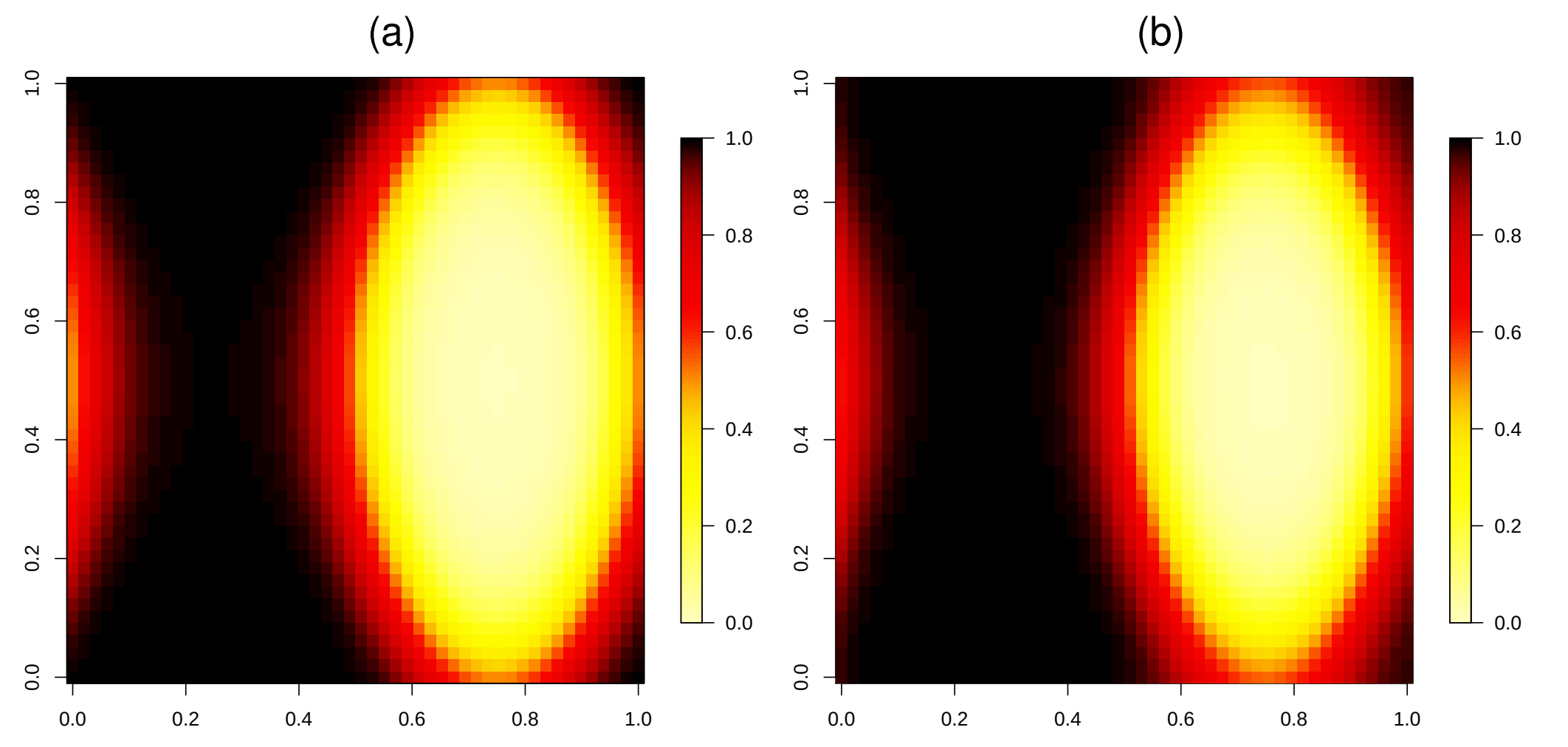

Figure 1. Maps of the theoretical (a) and estimated (b) probabilities of exceeding a threshold of 2.5.

- Additionally, the bootstrap procedure was also applied using the true covariance matrices $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_{\hat{\varepsilon}}$ (to examine the effect due to variogram estimation) and only considering the uncorrected residual-based variogram estimator.

- A summary of the squared errors, for $\sigma^2 = 0.16$, $r = 0.5$ and $c_0 = 0.04$, is shown in Table 1, where it is observed the good performance of the proposed procedure (these results could be compared with those in [6]). A similar behavior was observed in other simulation settings.

Table 1. Summary of squared errors ($\times 10^{-2}$) of the estimated probabilities.

	$n = 10 \times 10$			$n = 17 \times 17$			$n = 20 \times 20$		
	mean	median	sd	mean	median	sd	mean	median	sd
Theoretical	2.30	0.09	5.40	2.05	0.08	4.97	1.90	0.08	4.65
Residuals	5.00	0.16	11.00	4.10	0.16	9.40	3.80	0.16	8.60
Corrected	2.60	0.09	6.40	2.40	0.09	5.80	2.20	0.08	5.40

Application to real data

- The proposed methodology was applied to topsoil zinc (ppm) concentrations of a floodplain along the Meuse river near Stein, Netherlands [1]. This data set is supplied with the `gstat` package for `R`.

- Figure 2 shows the estimated trend function, the bias-corrected local linear variogram estimates and the fitted variogram model. In Figure 2(b), it is also shown the variogram model obtained using uncorrected residual-based variogram estimates. Note that, considering these estimates, it would be reasonable to (erroneously) assume independent errors and that the spatial variability of the data is captured by the trend.

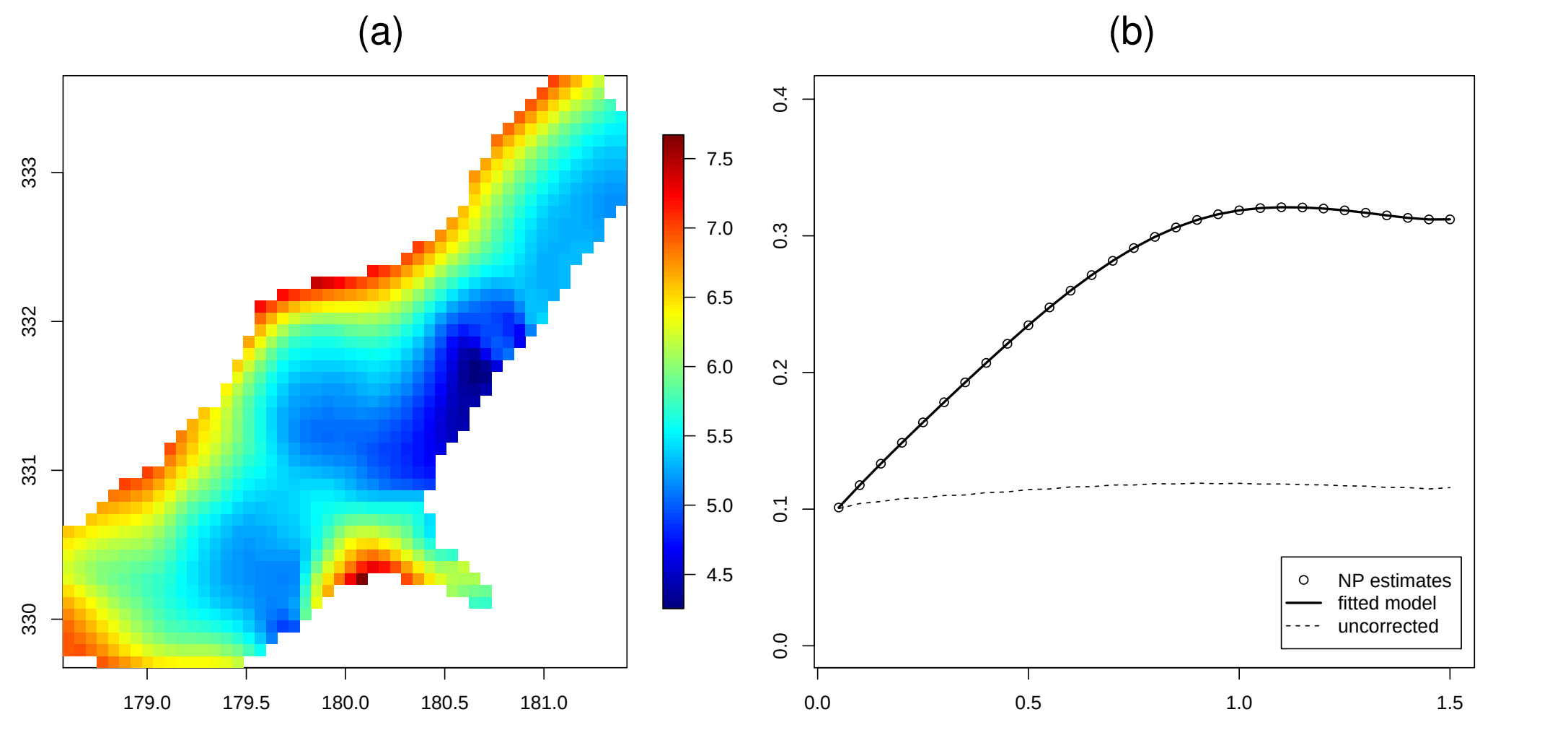

Figure 2. Map of nonparametric trend estimates (a), nonparametric bias-corrected pilot semivariogram estimates and fitted models (b) for log(zinc).

- Applying the bootstrap algorithm described above, estimated probability maps for several critical values were computed. For instance, Figure 3 shows the estimated conditional probabilities of log(zinc) equal or greater than $c = 6.0$.

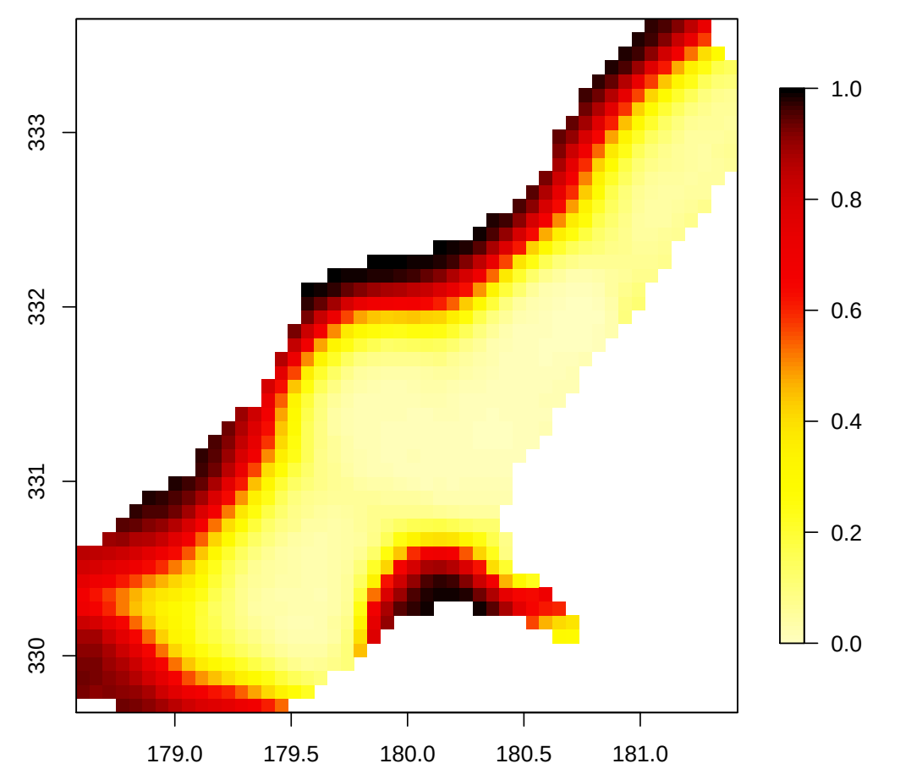

Figure 3. Estimated probability map for log(zinc) and $c = 6.0$.

Conclusions

- As observed in the simulation results, the proposed methodology yields accurate estimates of the risk probabilities.

- As the approach is fully nonparametric, problems due to model misspecification are avoided.

- The simulation results also confirm that the bias due to the direct use of residuals in variogram estimation may have a significant impact on risk assessment.

- This approach can be easily adapted to the construction of confidence or prediction intervals and to hypothesis testing.

- The procedure was implemented in the statistical environment `R`, using the functions for nonparametric trend and variogram estimation supplied with the `npsp` package (available on CRAN).

Acknowledgments

The research of Rubén Fernández-Casal and Mario Francisco-Fernández has been economically supported in part by the Consellería de Cultura, Educación e Ordenación Universitaria of the Xunta de Galicia through the agreement for the Singular Research Center CITIC. Additionally, the research of Mario Francisco-Fernández has been partially supported by Grant MTM2011-22392.

References

[1] Burrough, P.A. and McDonnell, R.A. (1998). *Principles of Geographical Information Systems*. Oxford University Press.

[2] Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, 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., González Manteiga, W. and Febrero-Bande, M. (2003). Flexible Spatio-Temporal Stationary Variogram Models. *Statistics and Computing* **13**, 127–136.

[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] Francisco-Fernández, M., Quintela-del Río, A. and Fernández-Casal, R. (2011). Nonparametric methods for spatial regression. An application to seismic events. *Environmetrics* **23**, 85–93.

[7] 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.

[8] Opsomer, J. D., Wang, Y. and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science* **16**, 134–153.

[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.