Special Issue Paper Environmetrics

Received: 4 January 2011, Revised: 24 October 2011, Accepted: 14 November 2011, Published online in Wiley Online Library: 19 December 2011

(wileyonlinelibrary.com) DOI: 10.1002/env.1146

Nonparametric methods for spatial regression. An application to seismic events†

Mario Francisco-Fernández[a], Alejandro Quintela-del-Río[a]* and Rubén Fernández-Casal[a]

Nonparametric regression estimation is a powerful tool to handle multidimensional data. When a dependent data set is analyzed, classical techniques need to be modified to provide useful results. In this work, different approximations to take the spatial dependence into account are exposed. A bandwidth selection technique that adjusts the generalized cross-validation criterion for the effect of spatial correlation, in the case of bivariate local polynomial regression, is considered. Moreover, a bootstrap algorithm is designed to assess the variability of the estimated spatial maps, and also to estimate the probability of obtaining a response variable larger than or equal to a given threshold, for a specific point. A simulation study checks the validity of the presented approaches in practice. The broad applicability of the procedures is demonstrated on a data set of earthquakes in the Iberian Peninsula. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: earthquakes; local polynomial regression; bootstrap; geostatistics

1. INTRODUCTION

Seismic series have been widely investigated using stochastic methods. The main purpose is to better understand the physic structure underlying the occurrence process of earthquakes. The most usual statistical models are of parametric type, assuming some specific distribution (Poisson, exponential) for the variables involved in the dynamics of the processes. But the parametric fits are not always good (see, e.g., Kijko and Graham, 1998). Moreover, parametric models can be affected by irregular events, because they usually assume a conventional seismic activity. On the contrary, nonparametric methods suppose a good alternative in the statistical analysis of earthquake data. They constitute a nice approach because of their ability to adapt to local variations, and they offer us the possibility of modeling the data without previous assumptions on particular theoretical models. Several researchers have performed earthquake analysis using nonparametric methods (see, e.g., Quintela-del-Río, 2010 and references therein).

In a general case, a seismic series is a spatiotemporal data set of observations of stochastic variables $\left(X_i^1, X_i^2, X_i^3, t_i, Y_i\right)$, where X_i^1 and X_i^2 are the latitude and longitude, respectively, of the epicenter, X_i^3 the depth, t_i the origin time and Y_i the magnitude. The present paper focuses on the estimation of the spatial trend of the earthquake magnitudes, by means of the model:

$$Y_i = m\left(\mathbf{X}_i\right) + \varepsilon_i \quad i = 1, 2, \ldots, n \tag{1}$$

where $m(\cdot)$ is the regression function, for which no specific parametric model is assumed, $\mathbf{X}_i = \left(X_i^1, X_i^2, X_i^3\right)$ refers to the epicenters locations and ε_i are random errors that may or may not be spatially correlated (no parametric distribution is assumed either).

The interest of this work is twofold. First, the estimation of the spatial pattern of earthquake magnitudes, with the aim of providing some information on the physics of the earthquake sizes. This will be carried out using nonparametric estimates of the function $m(\cdot)$. Specifically, the nonparametric local linear regression estimator (Wand and Jones, 1995) will be considered in the present paper. As it is well-known, this and other nonparametric estimators depend heavily on the bandwidth parameter, and a reliable automatic data selection procedure is needed to ensure good estimation results. The "bias-corrected and estimated" generalized cross-validation (GCV) criterion, proposed in Francisco-Fernández and Opsomer (2005) will be used here. The second interest of the current paper is the identification of areas with high and low probability of observing magnitudes larger than a given threshold, for events that occurred in that area. A bootstrap algorithm that takes into account the (possible) correlation in the data, combined with the previous nonparametric regression estimation techniques, will be used for this purpose. The reliability of the proposed methods will be checked for finite sample performance through a simulation study. These techniques will be also applied to a data catalog of earthquakes of the Northwest part of the Iberian Peninsula.

* *Correspondence to: Alejandro Quintela-del-Río, Depaamento de Matemáticas, Facultad de Informática, Campus de Elviña s/n, 15071 A Coruña, Spain. E-mail: aquintela@udc.es*

a *Universidad de A Coruña*

† *This article is published in Environmetrics as a special issue on Spatio-Temporal Stochastic Modelling (METMAV), edited by Wenceslao González-Manteiga and Rosa M. Crujeiras, University of Santiago de Compostela, Spain.*

The organization of this paper is the following: Section 2 presents the statistical model, reviews the nonparametric estimator under spatial dependence and describes the bootstrap method. In Section 3, simulation experiments evaluating the practical performance of our approach are presented. Finally, Section 4 illustrates the applicability of the methodology on a real earthquake data set.

2. STATISTICAL METHODS

Nonparametric regression estimation techniques can be applied to models with any finite number of explicative variables (of course, taking into account the so-called "curse of dimensionality" problem). Nevertheless, only a two-dimensional scheme (latitude, longitude) will be considered here. This is assumed for the sake of simplicity in the presentation and because, several times, the depth component for each recorded earthquake is missed in the data catalogs (this is, for instance, the case of the example presented in Section 4). However, note that all the presented techniques can be implemented, without loss of generality, with a larger number of regressor variables. Moreover, they could be extended taking into account additional spatiotemporal dimensions (including, for instance, the depth or even the temporal component) only through the errors ε_i, considering a more complicated spatiotemporal dependence (avoiding the curse of dimensionality in nonparametric estimation).

Assume that a set of $\mathbb{R}^3$-valued random vectors $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, is observed in a specific time interval, where Y_i are scalar response variables (magnitudes) and $\mathbf{X}_i$ are predictor variables (latitude and longitude of epicenter locations) with a common density f and compact support $\Omega \subseteq \mathbb{R}^2$. The relationship between the variables takes the form (1), where we suppose that $m(\cdot)$ is a smooth and continuous function and ε is a zero mean second-order stationary process with

$$\operatorname{cov}\left(\varepsilon_i, \varepsilon_j \,|\, \mathbf{X}_i, \mathbf{X}_j\right) = C(\mathbf{X}_i - \mathbf{X}_j) \tag{2}$$

where $C(\mathbf{u})$ is a positive-definite function, called the covariogram (with $C(\mathbf{0}) = \operatorname{var}(\varepsilon_i \,|\, \mathbf{X}_i) = \sigma^2$). If $C(\mathbf{u}) \equiv \sigma^2 \mathcal{I}_{\{\mathbf{0}\}}(\mathbf{u})$, where $\mathcal{I}_{\{\mathbf{0}\}}$ is the indicator function of the origin, the errors are independent. Other covariance functions establish the kind of spatial dependence. The estimation of the spatial trend $m(\cdot)$ is carried out by means of a nonparametric estimator with an automatic bandwidth selection procedure.

2.1. Local linear regression for spatial data

The classical local linear estimator, in the spatial framework, for $m(\cdot)$ at a location $\mathbf{x}$ is the solution for α to the least squares minimization problem

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{Y_i - \alpha - \boldsymbol{\beta}^T (\mathbf{X}_i - \mathbf{x})\right\}^2 K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})$$

where $\mathbf{H}$ is a 2×2 symmetric positive definite matrix, K is a bivariate kernel and $K_{\mathbf{H}}(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$. The well-known Epanechnikov kernel function $K(\mathbf{x}) = \frac{2}{\pi} \max\left\{\left(1 - \|\mathbf{x}\|^2\right), 0\right\}$ is used in this work. The bandwidth matrix $\boldsymbol{H}$ controls the shape and size of the local neighborhood used for estimating $m(\boldsymbol{x})$. Geometrically, when the kernel function is spherically symmetric with bounded support, this neighborhood is an ellipsis in $\mathbb{R}^2$ centered at $\mathbf{x}$. So, if the area of this ellipsis is small, we will obtain an undersmoothed estimation, with high variability; and, on the other hand, if the area is big, the resulting estimator will be very smooth and possibly with larger bias. The local linear regression estimator can be written explicitly as follows:

$$\hat{m}_{\mathbf{H}}(\mathbf{x}) = \mathbf{e}_1^T \left(\mathbf{X}_{\mathbf{x}}^T \mathbf{W}_{\mathbf{x}} \mathbf{X}_{\mathbf{x}}\right)^{-1} \mathbf{X}_{\mathbf{x}}^T \mathbf{W}_{\mathbf{x}} \mathbf{Y} \equiv s_{\mathbf{x}}^T \mathbf{Y} \tag{3}$$

where $\mathbf{e}_1$ is a vector with 1 in the first entry and all other entries 0, $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{W}_{\mathbf{x}} = \mathtt{diag}\{K_{\mathbf{H}}(\mathbf{X}_1 - \mathbf{x}), \ldots, K_{\mathbf{H}}(\mathbf{X}_n - \mathbf{x})\}$, and $\mathbf{X}_{\mathbf{x}}$ is a matrix with ith row equal to $(1, (\mathbf{X}_i - \mathbf{x})^T)$.

The interested reader is referred to the monography of Wand and Jones (1995) for theoretical results and practical applications of the local linear regression estimator in different settings. Correlated data can be also treated with this kind of estimators. Opsomer *et al.* (2001) provides an excellent review in the univariate correlated context. In a spatial dependence framework, Francisco-Fernández and Opsomer (2005) studied the bandwidth selection problem. They developed a GCV-based method that allows for the presence of correlated errors.

In contrast to the use of the estimator (3) with an adjusted bandwidth, other methods explicitly incorporate a correlation structure in the trend estimator. In time series context, different modifications of standard local polynomial regression have been proposed (e.g., Vilar-Fernández and Francisco-Fernández, 2002 or Xiao *et al.*, 2003). The basic idea behind these methods consists in transforming the original model to obtain an uncorrelated residual term. Applying this idea to model (1), the expression of the obtained estimator is

$$\tilde{m}_{\mathbf{H}}(\mathbf{x}) = e_1^T \left(\mathbf{X}_{\mathbf{x}}^T {\mathbf{P}^{-1}}^T \mathbf{W}_{\mathbf{x}} \mathbf{P}^{-1} \mathbf{X}_{\mathbf{x}}\right)^{-1} \mathbf{X}_{\mathbf{x}}^T {\mathbf{P}^{-1}}^T \mathbf{W}_{\mathbf{x}} \mathbf{P}^{-1} \mathbf{Y}$$

where $\mathbf{P}$ is the square root of the variance–covariance matrix $\boldsymbol{\Sigma} = \mathbf{P}\mathbf{P}^T$ of $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Note that in most cases, the matrix $\boldsymbol{\Sigma}$ is unknown and should be estimated in practice, using for example, one of the methods described in Section 2.3. This approach was analyzed in Vilar-Fernández and Francisco-Fernández (2002) for the univariate case, assuming an AR(1) correlation structure (where an explicit expression for matrix $\mathbf{P}^{-1}$ is available). These authors proved that under increasing domain asymptotics, the first-order properties of the estimator are not improved, although they showed the better finite sample behavior in terms of mean squared error (MSE), through simulations. A

similar approach was proposed by Høst (1999) for the spatial case, but the transformation to independent data is (locally) carried out after weighting the observations. Any of these approximations turns out to be interesting if the final aim is only the trend estimation. Nevertheless, because we are also interested in the distribution analysis of the estimates by bootstrap techniques, the high computing time needed for large sample sizes makes these estimators unhelpful in the spatial case.

2.2. Bandwidth selection

Under independence assumptions, classic smoothing parameters criteria for nonparametric regression are, for example, of cross-validation type (Craven and Whaba, 1979). These methods consider selecting the bandwidth $\mathbf{H}$ that minimizes the GCV function:

$$\mathrm{GCV}(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)}{1 - \frac{1}{n}\mathrm{tr}(\mathbf{S})}\right)^2 \tag{4}$$

with $\mathbf{S}$ is the $n \times n$ matrix whose ith row is equal to $\mathbf{s}_{\mathbf{X}_i}^T$, the smoother vector for $\mathbf{x} = \mathbf{X}_i$, and tr($\mathbf{S}$) its corresponding trace. Finding the minimizer of this function over the $d(d+1)/2$ ($d = 2$ when only longitude and latitude are considered) parameters in $\mathbf{H}$ can be achieved using numerical algorithms, as implemented in statistical software. However, this criterion should not be used directly for bandwidth selection with dependent errors, because its expectation is severely affected by the correlation (Liu, 2001). In this case, we suggest the use of the " bias-corrected" GCV criterion, proposed in Francisco-Fernández and Opsomer (2005), based on selecting the bandwidth $\mathbf{H}$ that minimizes the function

$$\mathrm{GCV}_c(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)}{1 - \frac{1}{n}\mathrm{tr}\,(\mathbf{SR})}\right)^2 \tag{5}$$

with $\mathbf{R}$, the correlation matrix of the errors. In practice, matrix $\mathbf{R}$ is unknown and it has to be estimated. Although Francisco-Fernández and Opsomer (2005) used estimators of the correlation matrix based on the method of the moments, other techniques could be employed for this task. In Section 2.3, different methods for estimating the covariogram function (2) will be presented. Once this function is estimated, it is easy to obtain an estimator of $\mathbf{R}$. The approach of Francisco-Fernández and Opsomer (2005) to select the bandwidth matrix in this context will be used here, but slightly modified using these estimators of the correlation matrix. We refer to Francisco-Fernández and Opsomer (2005), where the theoretical and practical optimality properties of this criterion were discussed, for a general description of the algorithm to select the bandwidth $\mathbf{H}$.

2.3. Spatial dependence modeling

In traditional geostatistical methods, the problem of modeling the spatial dependence between the data is a critical one. The characterization of the spatial dependence is usually carried out through the variogram $\gamma(\mathbf{u}) = C(\mathbf{0}) - C(\mathbf{u})$ (see, e.g., Cressie, 1993, Section 2.4.1, for an explanation on why the variogram estimation is preferred to the covariogram estimation). A classical approach consists in fitting a parametric model to a set of empirical semivariogram estimates. For instance, one of the best known isotropic variogram families is the Matern class (e.g., Stein, 1999, pp. 48–52):

$$\gamma_\theta(\mathbf{u}) = c_0 + c_1\left(1 - \frac{\left(\frac{\|\mathbf{u}\|}{a}\right)^\nu \mathcal{K}_\nu\left(\frac{\|\mathbf{u}\|}{a}\right)}{2^{\nu-1}\Gamma(\nu)}\right) \tag{6}$$

for $\mathbf{u} \neq \mathbf{0}$ ($\gamma_\theta(\mathbf{0}) = 0$), where $\mathcal{K}_\nu$ is the modified Bessel function of the second type of order ν, c_0 is the nugget effect, c_1 is the partial sill (the variance $\sigma^2 = c_0 + c_1 = C(\mathbf{0})$ is also called sill in this context), a is a scale parameter (proportional to the autocorrelation range) and ν is a smoothness parameter (which determines the shape of the covariogram at small lags). The case of $\nu = 0.5$ corresponds to the exponential model. Note also that $c_1 = 0$ is equivalent to independence.

In the spatial data context, under nonstationarity of the mean, the spatial dependence is typically modeled through the errors. Therefore, the variogram estimation requires a pilot estimate of the trend which, in turn, requires knowledge of the dependence. The algorithm mentioned in Section 2.2 circumvents this problem, analogously to the Neuman and Jacobson (1984) approach for the linear case. In this context, the classical variogram pilot estimator is

$$\hat{\gamma}(\mathbf{u}_l) = \frac{1}{2|N(\mathbf{u}_l)|}\sum_{N(\mathbf{u}_l)}\left(\hat{\varepsilon}_i - \hat{\varepsilon}_j\right)^2 \quad l = 1, \ldots, L \tag{7}$$

where L is the number of lags, $N(\mathbf{u}) = \{(i,j) : \mathbf{X}_i - \mathbf{X}_j \in \mathrm{tol}(\mathbf{u})\}$, tol($\mathbf{u}$) is a tolerance region around $\mathbf{u}$ and $|N(\mathbf{u})|$ denotes the number of contributing pairs at lag $\mathbf{u}$. Generally, the lags $\mathbf{u}_l$ are chosen regularly spaced.

To obtain the parameter estimates $\hat{\theta}$, one of the most widely used fitting criteria (see e.g., Cressie, 1993, Section 2.6) is the weighted least squares (WLS) method:

$$\hat{\theta} = \arg\min_\theta \sum_{l=1}^{L} \omega_l \left(\gamma_\theta(\mathbf{u}_l) - \hat{\gamma}(\mathbf{u}_l)\right)^2$$

Following the recommendations of Journel and Huijbregts (1978, p. 74), the fit of a valid model is usually carried out up to the half of the maximum possible lag and considering only pilot estimations with at least 30 contributing pairs (i.e., $|N(\mathbf{u})| \geq 30$). This gives a practical rule for selecting L and the tolerance regions. The weights ω_l are usually chosen following the idea proposed by Cressie (1985). However, we applied an iterative procedure, with $w_l^{(0)} = 1$ (ordinary least squares) used for the first step and with the weights recalculated for each iteration until convergence (i.e., $\omega_l^{(k)} = |N(\mathbf{u})|/\gamma_{\hat{\theta}^{(k-1)}}(\mathbf{u}_l)$ at iteration k). Specifically, the parametric fits considered in the following sections were obtained combining weighted least squares linear regression with a modified Levenberg-Marquardt algorithm (see. e.g. Fernández-Casal *et al.*, 2003).

Nevertheless, in practice, a parametric variogram fit to the empirical variogram could be unsatisfactory. For instance, the assumption of isotropy (or geometric anisotropy) could be not appropriate. Fernández-Casal *et al.* (2003) proposed the modeling of anisotropies through the combination of flexible models with nonparametric semivariogram pilot estimation, avoiding problems related to misspecification of the variogram model, and obtaining more efficient estimates. The anisotropic two component variogram models proposed by these authors, called extended Shapiro-Botha (Shapiro-Botha, 1991) models (denoted by SVSBE(d_1, d_2)), are of the form:

$$\gamma(\mathbf{u}) = \nu_0 - \sum_{k=1}^{K} \kappa_{d_1}(x_{1_k} |u_1|)\kappa_{d_2}(x_{2_k} |u_2|)z_k \tag{8}$$

where $\mathbf{x}_k = \left(x_{1_k}, x_{2_k}\right)$ are user defined discretization points,

$$\kappa_d(x) = \left(\frac{2}{x}\right)^{(d-2)/2} \Gamma\left(\frac{d}{2}\right) J_{(d-2)/2}(x), \quad \kappa_\infty(x) \equiv e^{-x^2}$$

J_p is the Bessel function of order p and the $(K+1)$-dimensional vector of parameters $(z_1, \ldots, z_K, \nu_0)$ verifies the linear restrictions:

$z_{kl} \geq 0, k = 1, \ldots, K$; and

$$c_0 = \nu_0 - \sum_{k=1}^{K} z_k \geq 0$$

If the WLS criterion is chosen, the fitting of such models to pilot estimates of the variogram can be easily carried out in practice by means of quadratic programming.

The models obtained in this way are extremely flexible, and, in general, they always produce good approximations to the pilot estimates. To avoid problems related with "overfitting", the use of a nonparametric variogram estimator is also proposed. Analogous to the case of trend estimation, the pilot local linear estimate $\hat{\gamma}(\mathbf{u})$ is the solution for α to the least squares minimization problem:

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i<j} \left\{\left(\hat{\varepsilon}_i - \hat{\varepsilon}_j\right)^2 - \alpha - \boldsymbol{\beta}^T \left(\mathbf{X}_i - \mathbf{X}_j - \mathbf{u}\right)\right\}^2 K_{\mathbf{G}}\left(\mathbf{X}_i - \mathbf{X}_j - \mathbf{u}\right) \tag{9}$$

where $\mathbf{G}$ is the corresponding bandwidth matrix. After considering different selection criteria, Fernández-Casal *et al.* (2003) recommended the use of the standard cross-validation method. Therefore, $\hat{\gamma}(\mathbf{u})$ has a similar expression to $\hat{m}_{\mathbf{H}}(\mathbf{x})$ in (3), but now the response variable $\mathbf{Y}$ is a vector of length $n^* = n(n-1)/2$ with components $\left(\hat{\varepsilon}_i - \hat{\varepsilon}_j\right)^2$, $1 \leq i < j \leq n$, and $\mathbf{X}_\mathbf{x}$ is a $(n^* \times 2)$ matrix with ith row equal to $(1, \left(\mathbf{X}_i - \mathbf{X}_j - \mathbf{u}\right)^T)$

Note also that the variogram estimate based on the residuals induces a bias in the estimation of the dependence parameters. However, following Cressie (1993, pp. 296–299), the influence of this bias in the trend estimation is expected to be small (although it may lead to underestimation of the variance). We performed a small simulation experiment to evaluate this effect, and an underestimation of the spatial variability was also observed. A modification of the dependence estimation process could be applied to correct this bias (for instance, using a similar procedure to the one proposed by Beckers and Bogaert, 1998, for the linear case). However, we did not explore this approach further, because we have observed that this underestimation feature does not have an important impact on the bandwidth selection problem and, therefore, on the estimation of the spatial trend and the general bootstrap approach described later.

Additionally, traditional Kriging models typically assume that the response variable is a simpler (linear or constant) function of locations and a spatially correlated noise. Therefore, in these models, most of the observed behavior of the data is attributed to noise and not to the underlying mean function. Consequently, the error term should be carefully modeled (even considering a complicated spatial dependence structure). This fact is in contrast with model (1), which assumes that the response variable Y is the sum of two components, an unknown smooth function of locations, and a zero-mean (second-order stationary) error term. With a model like (1), assumptions about the error term are expected to have less effect on the obtained conclusions. Moreover, note that it is not possible to determine the correct model for the data without replication at the same locations, and in practice, different approaches could lead to similar estimated (or predicted) spatial maps (for more details see, for instance, Altman, 1997, or Cressie, 1993, Section 3.1).

2.4. Bootstrap algorithm

The nonparametric regression methods described in the previous sections are combined with a bootstrap procedure to evaluate the accuracy of the nonparametric estimates, and to estimate the probability of obtaining a response variable in a specific location exceeding a certain

threshold. The algorithm used (detailed later) extends to the spatial context the parametric bootstrap of Vilar-Fernández and González-Manteiga (1996), who also considered a dependence situation. The specific steps are the following:

1. Obtain, using the algorithm in Section 2.2, the optimal bandwidth matrix $\mathbf{H}$, the residuals $\hat{\varepsilon}_i = Y_i - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)$, $i = 1, \ldots, n$, and the estimated covariance function $\hat{C}$.
2. Generate bootstrap samples with the estimated spatial trend $\hat{m}_{\mathbf{H}}(\mathbf{X}_i)$ and adding bootstrap errors generated as a spatially correlated set of errors. The bootstrap errors are obtained as follows:
   (a) Using the estimated covariogram $\hat{C}$, compute the variance–covariance matrix of the residuals $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n)$, denoted by $\hat{\mathbf{V}}$.
   (b) Find the matrix $\mathbf{P}$, such that $\hat{\mathbf{V}} = \mathbf{P}\mathbf{P}^T$, using Cholesky decomposition.
   (c) Compute the "independent" variables $\mathbf{e} = (e_1, e_2, \ldots, e_n)$, given by $\mathbf{e} = \mathbf{P}^{-1}\hat{\boldsymbol{\varepsilon}}$.
   (d) These independent variables are centered and, from them, we obtain an independent bootstrap sample of size n, denoted by $\mathbf{e}^* = (e_1^*, e_2^*, \ldots, e_n^*)$.
   (e) Finally, the bootstrap errors $\hat{\boldsymbol{\varepsilon}}^* = (\hat{\varepsilon}_1^*, \ldots, \hat{\varepsilon}_n^*)$ are $\hat{\boldsymbol{\varepsilon}}^* = \mathbf{P}\mathbf{e}^*$, and the bootstrap samples are $Y_i^* = \hat{m}_{\mathbf{H}}(\mathbf{X}_i) + \hat{\varepsilon}_i^*$, $i = 1, 2, \ldots, n$.
3. The nonparametric local linear regression estimator is applied for each bootstrap sample, using the same bandwidth $\mathbf{H}$ as for the original analysis, and a map of probability areas (magnitude larger than or equal to a threshold) is produced. This process is repeated a large number of times B (in our analysis, $B = 1000$). In a final point, a map with the frequencies (across bootstrap replicates and for each location) of how often that location is included in the at-risk area is computed.

The spatial correlation for the errors is taken into account in the previous algorithm. If the errors were independent, this procedure can be readily shortened by computing the residuals $(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)$ with the classical GCV method for bandwidth selection in the first step, and next, obtaining the bootstrap errors $(\hat{\varepsilon}_1^*, \ldots, \hat{\varepsilon}_n^*)$ by resampling with replacement among the original residuals.

3. SIMULATION EXPERIMENTS

In this section, the behavior of the previous approach is evaluated in a simulation study. The main goals of the simulation are to compare the performance of the methods when the parametric and the nonparametric estimators of the variograms (described in Section 2.3) are used, and to evaluate the effect of different model specifications, sample sizes and dependence structures. For this purpose, $N = 1000$ samples of size n are generated following the model (1), where the design points $\mathbf{X}_i = (X_i^1, X_i^2)$ are uniformly sampled in the unit rectangle, the mean function is the additive model $m(x, y) = 2.5 + \sin(2\pi x) + 4(y - 0.5)^2$, and the random errors ε_i are normally distributed with zero mean and exponential covariance function (see Equation (6)). In our experiments, the sill or variance was fixed to $\sigma^2 = 0.16$ ($c_1 = \sigma^2 - c_0$). Two spatial correlation degrees, corresponding to practical ranges of 0.5 (strong correlation) and 0.15 (moderate correlation), and three nugget values of $c_0 = 0$, 0.04 and 0.08 (100%, 75% and 50% of spatial variability, respectively) are considered. Sample sizes of $n = 200$, 300 and 400 are used in the study.

Given a threshold u, and using the model employed to generate the samples, it is easy to obtain the theoretical probabilities of obtaining a value of the response variable larger than u in any point of the region of study. These theoretical probabilities are computed in a regular 100×100 grid. Next, our complete procedure is applied over the same grid (with $B = 1000$ bootstrap replications) and N estimated probability maps are computed with the generated samples. Finally, bias, variance and mean square error maps, computed in each one of the grid points, are obtained. Moreover, by averaging over the grid points, the average bias (AB), the average variance (AV) and the average mean square error (AMSE) values are calculated. The "bias-corrected and estimated" GCV bandwidth selection method and the bootstrap algorithm proposed in Section 2.4 require the estimation of the variance–covariance matrix of the errors. Two different models are used in our simulations for that purpose, an isotropic exponential and a more flexible SVSBE model (fitted in each case as explained in Section 2.3). Additionally, the true variance–covariance matrix is also used in our experiments. Obviously, using the true covariance matrix is an unrealistic setting, but it is included in the simulations as a benchmark.

For the sake of brevity, only results with $n = 200$ and a threshold value of $u = 3.0$ are shown here. Complete simulation results are available in the accompanying Supporting Materials[‡]. The AB, AV and AMSE values ($\times 10^{-2}$) are presented in Table 1 for $c_0 = 0$ and $r = 0.5, 0.15$, and in Table 2 for $c_0 = 0.04, 0.08$ and $r = 0.5$.

It can be observed in Tables 1 and 2 that, as expected, the best results are obtained when the true covariance matrix is used in the procedure. It is also observed in both tables that decreasing the spatial dependence improves the behavior of the estimators. On the other hand, using the parametric exponential model or the nonparametric estimator of the covariogram produced very similar results. Note that the SVSBE modeling gives very good results, with a very similar performance to the parametric fit when the correlation is strong and even outperforming when the correlation is weak (this may be due to the better performance of the local linear pilot semivariogram estimator). The good performance of the proposed approach can be observed in Figure 1, where it is shown in the left panel, the theoretical probabilities of obtaining a value of the response variable larger than 3.0 in the region of study, and in the right panel, the corresponding average estimated probabilities over the N replicas, when the SVSBE model is used. A similar plot is obtained when the parametric exponential dependence structure is assumed.

The simulation study is completed in several directions (see the Supporting Materials). For example, we considered the situation when the parametric covariance function is misspecified or when the design points $\mathbf{X}_i$ are not uniformly sampled. Similar results as those shown in Table 1 and Figure 1 are obtained in both cases, and even in situations when the covariance model cannot be specified exactly, our approach assuming an exponential model provided good results. We expect our approach using an isotropic exponential covariogram model to be robust to model misspecification, as long as the correlation can be reasonably well approximated by a positive and decreasing

[‡]Supporting information may be found in the online version of this article.

Table 1. Average bias ($\times 10^{-2}$), average variance ($\times 10^{-2}$) and average MSE ($\times 10^{-2}$), when $n = 200$, $u = 3.0$, $c_0 = 0$ and practical ranges of $r = 0.5$ and $r = 0.15$.

Covariance	$r = 0.5$			$r = 0.15$		
estimator	AB	AV	AMSE	AB	AV	AMSE
True	4.23	2.83	3.38	4.03	2.40	2.64
Exponential	4.41	4.40	4.96	4.03	3.31	3.55
SVSBE	4.39	4.47	5.03	4.00	2.91	3.15

Note: MSE, mean squared error; AB, average bias; AV, average variance; AMSE, average mean square error; SVSBE, Shapiro-Botha Variogram model

Table 2. Average bias ($\times 10^{-2}$), average variance ($\times 10^{-2}$) and average MSE ($\times 10^{-2}$), when $n = 200$, $u = 3.0$, $r = 0.5$ and nugget values of $c_0 = 0.04$ and $c_0 = 0.08$ ($c_1 = \sigma^2 - c_0$; $\sigma^2 = 0.16$).

Covariance	$c_0 = 0.04$			$c_0 = 0.08$		
estimator	AB	AV	AMSE	AB	AV	AMSE
True	3.69	2.59	2.99	2.95	2.39	2.66
Exponential	3.82	3.83	4.23	3.06	3.46	3.74
SVSBE	3.80	3.98	4.39	3.05	3.45	3.72

Note: MSE, mean squared error; AB, average bias; AV, average variance; AMSE, average mean square error; SVSBE, Shapiro-Botha Variogram model

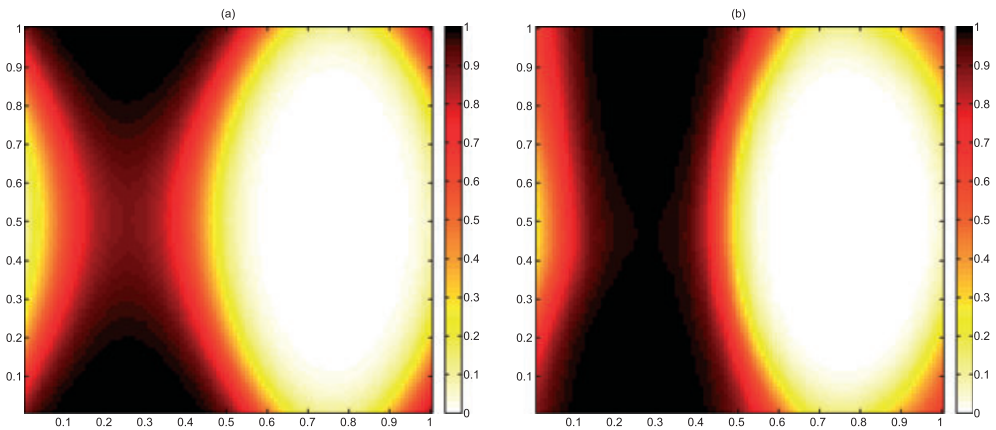

Figure 1. Probabilities of exceeding threshold 3.0 considering the exponential model with $c_0 = 0$ and $r = 0.5$ as theoretical covariogram: (a) Theoretical probabilities, and (b) mean estimated probabilities using the SVSBE model

function of distance between observations, as is often observed in practice. In other cases, the nonparametric modeling will possibly produce better results.

4. DATA ANALYSIS

In this section, the statistical procedures described in Section 2 are applied to a data set of earthquakes. The northwest (NW) part of the Iberian Peninsula has been considered. This area is limited by the coordinates 41°N–44°N and 6°W–10°W, which contains the autonomic region of Galicia (Spain) and northern Portugal. A detailed geophysical description of this zone can be seen in Francisco-Fernández and Quintela-del-Río (2011). The considered data set consists of $N = 1643$ earthquakes (with magnitude above or equal to 2.0 in Richter's scale), which occurred from 25 November 1944 to 03 April 2008. These were obtained from the National Geographic Institute (IGN) of Spain. It can be searched using the web page www.ign.es/ign/es/IGN/SisCatalogo.jsp. In Figure 2, a map of this region, together with the selected data set is displayed.

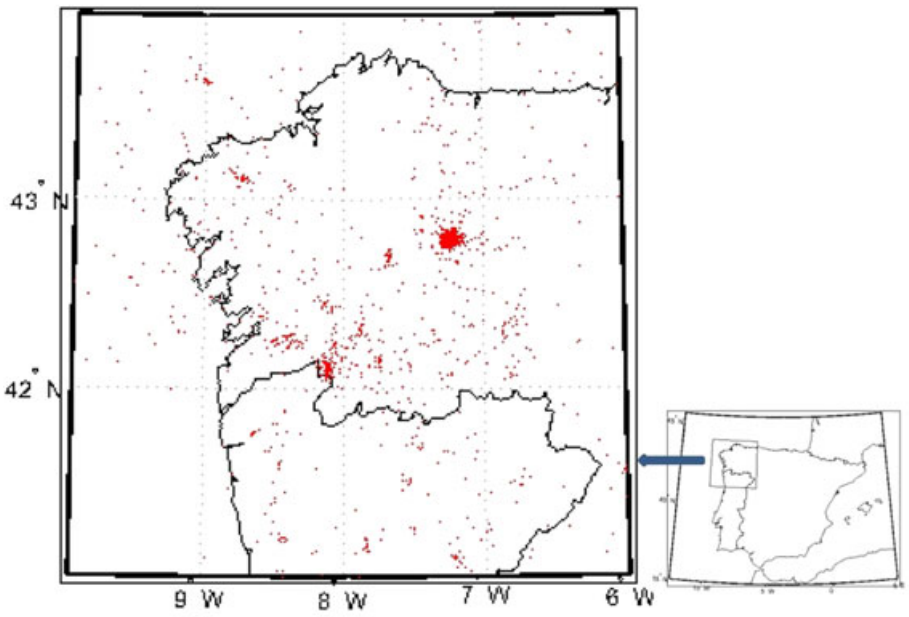

Figure 2. Earthquake locations

Using model (1), a nonparametric regression estimator (3) was fitted to the data set. To apply the adjusted GCV method for bandwidth selection of Francisco-Fernández and Opsomer (2005), it is necessary to estimate the error model correlation. The traditional geostatistical methodology described at the beginning of Section 2.3 was used in the first step. A preliminary analysis of the data, plotting directional sample variograms, suggests that an isotropic exponential model for the covariogram (Equation (6) with $\nu = 0.5$) describes well the spatial dependence of the residuals. The accompanying Supporting Materials contain a figure showing the pilot variogram estimates and the fitted variogram model. The estimated covariogram parameters in (6) were $\hat{c}_0 = 0.066$, $\hat{c}_1 = 0.132$ and $\hat{a} = 0.047$ (corresponding to a practical range of 0.141 in degrees), using $L = 20$ and $u_l = l \cdot 0.04$ in (7). Next, the flexible approach described at the end of Section 2.3 was also used, and an SVSBE(3,3) (see Equation (8)) was fitted by WLS. The local linear pilot semivariogram estimates and the fitted SVSBE model are shown in Figure 3.

The bandwidth matrices obtained with the selection criterion (5) were:

$$\mathbf{H}_{GCV1} = \begin{bmatrix} 0.578 & 0.000 \\ 0.000 & 1.127 \end{bmatrix}, \mathbf{H}_{GCV2} = \begin{bmatrix} 0.567 & 0.000 \\ 0.000 & 1.130 \end{bmatrix}$$

denoting by $\mathbf{H}_{GCV1}$ and $\mathbf{H}_{GCV2}$ the bandwidth matrices corresponding to the parametric and the flexible variogram models, respectively (the units of the values in the matrices are degrees). As expected, given the similarity of the obtained fits, both matrices are almost identical,

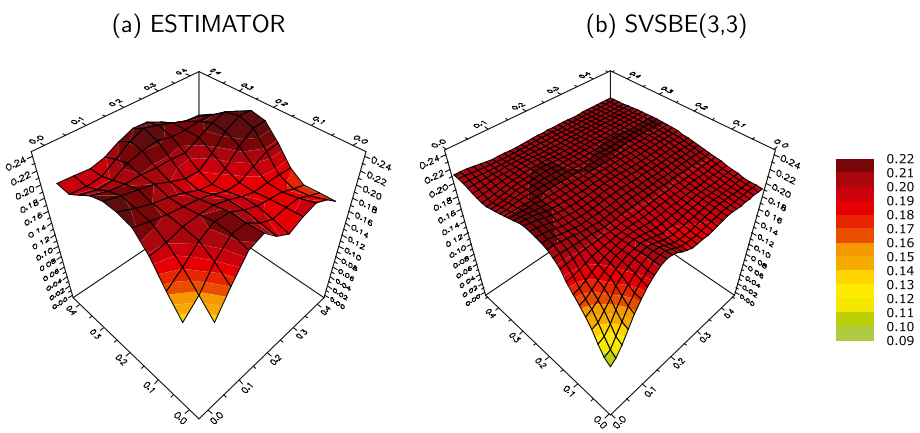

Figure 3. (a) Local linear semivariogram estimates. (b) Fitted extended Shapiro-Botha model

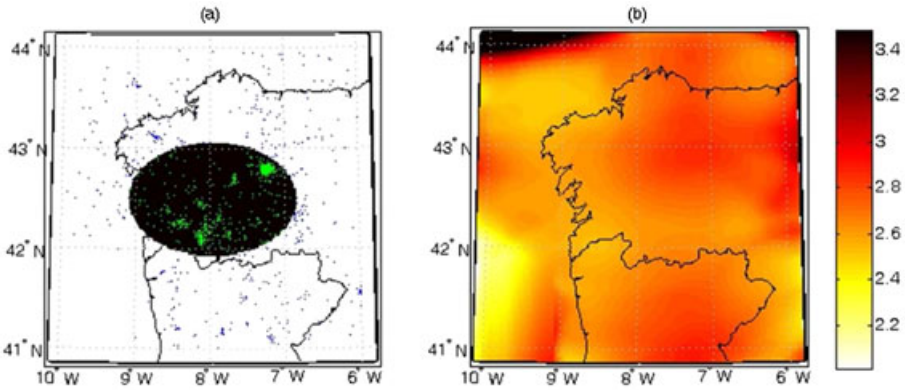

Figure 4. (a) Earthquake locations and shape of the optimal bandwidth at coordinates 42.5^{o} N and 8° W. (b) Estimated spatial trend

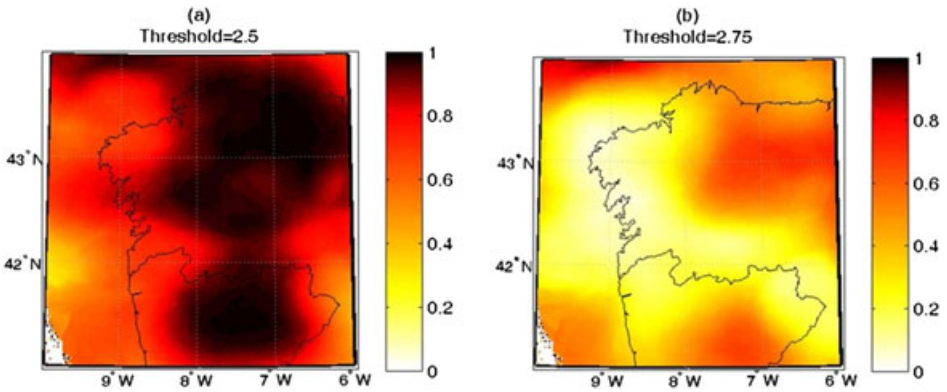

Figure 5. Maps with bootstrap probabilities of observing a magnitude larger than or equal to a threshold value, for earthquakes that occurred in the northwest of the Iberian Peninsula. (a) Threshold value of 2.5 and (b) threshold value of 2.75

and also other results obtained with both approaches (most of the spatial variability is captured by the trend estimator, as mentioned at the end of Section 2.3). Therefore, only the results obtained when using the previous estimated isotropic exponential model for the covariogram (and therefore, the bandwidth $\mathbf{H}_{GCV1}$) will be shown. Both the $\mathbf{H}_{GCV1}$ bandwidth (at coordinates 42.5°N and 8°W) and the corresponding regression estimator in the area of interest are shown in Figure 4 (the analogous plot using $\mathbf{H}_{GCV2}$ is almost identical). With this bandwidth, a percentage between 20% and 35% (approximately) of the observations have non-zero weight in the kernel function of the nonparametric regression fit, in any location $\mathbf{x}$. An important observation about the estimates shown in Figure 4 is that the extreme low and high estimated values occur at the boundary of the estimation region, and, although the use of local linear weights with a compactly supported kernel greatly minimizes the edge effects, these results should be carefully examined.

As it can be observed in Figure 4, seismicity is scattered, and three areas deserve to be mentioned: the centered at the Becerreá zone (limited by the coordinates 42.75°N–43.5°N and 6.5° W–8°W) with the greater magnitudes, some groups of events around the 42nd parallel (the boundary between Spain and Portugal), with magnitude not too high, and the central-western part of Galicia with moderate seismicity (low magnitude without high isolated episodes). The results agree with geological conclusions obtained before (Rueda and Mezcua, 2001; Estévez *et al*, 2002).

The pattern of probability of earthquake occurrences is now analyzed using the bootstrap procedure described in Section 2.4. In Figure 5 the pointwise bootstrap probabilities are mapped. They show that the probability of the magnitude of an earthquake is above or equal to a considered threshold (supposed that an earthquake occurs in a certain point of this region).

We considered two threshold values to evaluate the sensitivity to the choice of this parameter, 2.5 in the left picture and 2.75 in the right one. The differences are clear between both maps. In Figure 5(a), a large proportion of the area has a high probability of observing magnitudes 2.5 or above, for earthquakes occurring there, whereas in Figure 5(b), only the earthquakes occurring in the areas of Becerreá and northern Portugal have a high probability of having a magnitude 2.75 or greater. The discrepancies are mainly due to the fact that the values of the mean magnitude range essentially from 2 and 3, with no significant differences in the region. Thus, small changes in the threshold, from 2.5 to 2.75, can cause large differences in the plots.

Acknowledgements

The research of Mario Francisco-Fernández has been partially supported by Grants MTM2008-00166 (ERDF included) and MTM2011-22392. The research of Rubén Fernández-Casal has been partially supported by MEC Grant MTM2008-03010. The authors thank the associate editor and two referees for constructive comments that improved the presentation of this article.

REFERENCES

Altman NS. 1997. Krige, smooth, both or neither? In *ASA Proceedings of the Section on Statistics and the Environment*. American Statistical Association: Alexandria, VA; 60–65.

Beckers F, Bogaert P. 1998. Nonstationarity of the mean and unbiased variogram estimation: extension of the weighted least squares method. *Mathematical Geology* **30**: 223–240.

Craven P, Wahba G. 1979. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics* **31**: 377–403.

Cressie N. 1985. Fitting variogram models by weighted least squares. *Mathematical Geology* **17**: 563–586.

Cressie N. 1993. *Statistics for Spatial Data*. John Wiley & Sons: New York.

Estévez-Pérez G, Lorenzo-Cimadevila H, Quintela-del-Río A. 2002. Nonparametric analysis of the time structure of seismicity in a geographic region. *Annals of Geophysics-Italy* **45**: 497–512.

Fernández-Casal R, González-Manteiga W, Febrero-Bande M. 2003. Space–time dependency modeling using general classes of flexible stationary variogram models. *Journal of Geophysical Research* **108**: 8779.

Francisco-Fernández M, Opsomer JD. 2005. Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. *Canadian Journal of Statistics* **33**: 539–558.

Francisco-Fernández M, Quintela-del-Río A. 2011. Nonparametric seismic hazard estimation: a spatio–temporal application to the northwest of the Iberian Peninsula. *Tectonophysics* **505**: 35–43, DOI: 10.1016/j.tecto.2011.04.001.

Høst G. 1999. Kriging by local polynomials. *Computational Statistics and Data Analysis* **29**: 295–312.

Journel AG, Huijbregts CJ. 1978. *Mining Geostatistics: 22–47*. Academic Press: London.

Kijko A, Graham G. 1998. Parametric-history procedure for probabilistic seismic hazard analysis. Part I: estimation of maximum regional magnitude $m_{\max}$. *Pure and Applied Geophysics* **152**: 413–442.

Liu X. 2001. Kernel smoothing for spatially correlated data. *Ph. D. thesis*, Department of Statistics, Iowa State University.

Neuman SP, Jacobson EA. 1984. Analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels. *Mathematical Geology* **16**: 499–521.

Opsomer JD, Wang Y, Yang Y. 2001. Nonparametric regression with correlated errors. *Statistical Science* **16**: 134–153.

Quintela-del-Río A. 2010. On nonparametric techniques for area-characteristic seismic hazard parameters. *Geophysical Journal International* **180**: 339–346.

Rueda J, Mezcua J. 2001. Sismicidad, sismotectónica y peligrosidad sísmica en Galicia. *IGN Technical Publication* **35**.

Shapiro A, Botha JD. 1991. Variogram fitting with a general class of conditionally non-negative definite functions. *Computational Statistics and Data Analysis* **11**: 87–96.

Stein ML. 1999. *Interpolation of Spatial Data, Some Theory for Kriging*, Springer Series in Statistics. Springer-Verlag: New York.

Vilar-Fernández JM, González-Manteiga W. 1996. Bootstrap test of goodness of fit to a linear model when errors are correlated. *Communications in Statistics. Theory and Methods* **25**: 2925–2953.

Vilar-Fernandez JM, Francisco-Fernández M. 2002. Local polynomial regression smoothers with AR-error structure. *Test* **11**: 439–464.

Wand MP, Jones MC. 1995. *Kernel Smoothing*. Chapman and Hall: London.

Xiao Z, Linton OB, Carroll RJ, Mammen E. 2003. More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *Journal of the American Statistical Society* **98**: 980–992.

