A nonparametric analysis of the spatial distribution of earthquake magnitudes

Mario Francisco-Fernández
Universidad de A Coruña *

Alejandro Quintela-del-Río †
Universidad de A Coruña *

Rubén Fernández Casal
Universidad de A Coruña *

August 24, 2010

Abstract

This article describes an application of nonparametric local linear regression to study the spatial structure of the mean trend of earthquake magnitudes. If spatial correlation is suspected in a data set of earthquakes in a particular geographic area, the smoothing parameter needed to obtain the estimator of the mean magnitude will be computed using a corrected version

*Departamento de Matemáticas, Facultad de Informática, Campus de Elviña s/n, A Coruña, 15071, Spain.

†Corresponding author. Email: aquintela@udc.es

of a generalized cross-validation method. This procedure allows to take the spatial dependence into account to obtain better smoothing parameters. Additionally, a parametric bootstrap technique is used to quantify the variability of the spatial maps produced with the nonparametric estimation method and to generate maps to try to quantify the seismic hazard in the area considered. These techniques are applied to two different earthquake data sets: the historic catalogue of the northwest of the Iberian Peninsula and the earthquakes in California from January 1998 to April 2008.

Key Words: earthquakes; magnitude; local polynomial regression; nonparametric estimation; parametric bootstrap.

1 Introduction

A seismic series is a set of earthquakes occurring in a given period of time in a given area. Earthquakes of a seismic series are considered stochastic mathematical variables, belonging to a continuous space-time-energy medium with dimension 5 $(X_i^1, X_i^2, X_i^3, t_i, Y_i)$, where X_i^1 and X_i^2 are the latitude and longitude, respectively, of the epicenter, X_i^3 the depth of the focus, t_i the origin time and Y_i the magnitude. If the activity develops without abrupt changes, it is possible to know the structure that exists between the earthquakes (Torcal *et al.*, 1999). One of the purposes of the research performed in this field is to better understand the physical mechanisms controlling the occurrence and size of earthquakes. To that end, an effective approach is to study the behavior of the seismic series using stochastic methods to try

to specify the relation between the five variables mentioned.

There have been numerous studies of seismic series using stochastic methods. From the initial works of Udias and Rice (1975) or Vere-Jones (1978), several papers present extensive statistical analyses and mathematical modelizations of complete data sets. Some classic papers on parametric modeling of seismic series are those of Ogata (1988, 1998), Kagan (1990, 1999) or Vere-Jones (1992), and also the more recent works of Kagan and Jackson (2000), Daley and Vere-Jones (2003) or Kijko (2004). Traditionally, the distribution of the earthquake size is described by the classic Gutenberg-Richter (G-R) law (Gutenberg and Richter, 1944) or magnitude-frequency relation

$$\log N = a - bm, \tag{1}$$

where N is the number of earthquakes with magnitude above or equal to m. The G-R relation (1) is also the basis of traditional methods of probabilistic seismic hazard assessment. It can be transformed into an exponential distribution or into a more general Pareto (power-law) distribution for the scalar seismic moment M (Kagan and Jackson, 2000). In this context, the b-value is proportional to the inverse of the mean magnitude.

Additionally, several studies have revealed the spatial heterogeneity of seismicity parameters (e.g., Ogata and Katsura, 1993; Wiemer and Benoit, 1996; Wiemer and Wyss, 1997, 2002). The method traditionally used to study the spatial variations in the frequency-magnitude distribution is the mapping of the b-values of the G-R distribution using a gridding technique. For each node of the grid, considering the

nearest epicenters, a value for the magnitude of completeness is first determined and then the b-value in the frequency-magnitude relation is estimated (assuming independent and identically distributed data). The physical interpretation of the spatial variation of the b-value can be related, for instance, to material heterogeneity, stress distribution or previous earthquake activity. The variations in this parameter also have a significant impact on probabilistic seismic hazard assessment, so it is crucial to have a reliable procedure for its estimation.

The parametric models traditionally used in the analysis of seismic data, however, do not always fit those data well. For instance, the linear relation of the G-R law can be affected in different ways (Kijko and Graham, 1998), partly because parametric models are usually well suited only to a sequence of seismic events with similar causes. Moreover, parametric models can be insensitive to anomalous events, since models tend to be formulated through experience of relatively conventional seismic activity. Since about 1970, there has been extensive statistical literature on what has been called "nonparametric curve estimation" (beginning with Parzen, 1962; Nadaraya, 1964, or Watson and Leadbetter, 1964). This methodology is a flexible and potent tool used to describe the behavior of univariate and multivariate data sets, because it does not need the specification of a particular model to work with (such as the normal distribution, or a linear relation). The nonparametric statistical techniques are, in some cases, a supplement for parametric models. In other cases, they are a valid alternative to the classic parametric techniques, where it is presupposed that the curve has a certain functional form depending on some parameters

that must be estimated (see, for example, Wand and Jones, 1995, or Simonoff, 1996). Nonparametric methods have also been used in the statistical analysis of earthquake data. The papers of Choi and Hall (1999), Stock and Smith (2002) or Zhuang *et al.* (2002) consider nonparametric estimation of the intensity function. Regarding regression estimation, the work of Grillenzoni (2005) develops adaptive modelings for earthquake data in northern California. Failure rate or hazard estimation is analyzed in the paper of Estévez *et al.* (2002) for data from NW Iberian Peninsula (the same region considered here), and other European areas. More recent papers, using similar methodologies to those applied here, are Lasocki and Orlecka-Sikora (2008) and Marsan and Lengliné (2008).

The present paper focuses on a nonparametric model for the magnitude distribution. No particular form is a priori assumed for this variable, allowing in this way the existence of deviations from the G-R relation (1). Specifically, we are interested in mapping the (two-dimensional) spatial distribution of the earthquake magnitudes, by means of the model

$$Y_i = m(X_i^1, X_i^2) + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

where $m(\cdot)$ is a regression function, for which a specific parametric model is not assumed, and ε_i are random errors that may or may not be spatially correlated (no parametric distribution is assumed either). Let $\mathbf{X}_i = (X_i^1, X_i^2)$ refer to the epicenters (latitude and longitude, respectively, expressed in degrees) corresponding to Y_i (magnitude). The nonparametric estimation of the spatial trend $m(\cdot)$ can provide

useful insights on the physics of the process, similar to those that can be obtained representing the b-value in the traditional approach. It can be also used, in combination with a bootstrap algorithm, to quantify the seismic hazard, allowing for the estimation of the probability that an earthquake of a magnitude larger than or equal to a given threshold occurs in a particular area. Nevertheless, the procedure presented in this paper remains valid when the selected area and/or the selected catalogue of data depart from the usual assumptions in earthquake stochastic modeling.

The goal of this article is to show the utility of a particular type of nonparametric estimation method, named "local linear regression" (Fan and Gijbels, 1996), for the spatial statistical analysis of earthquake data. The representation of the magnitude as a spatial function, jointly with the use of bootstrap techniques, provides a useful "first step" in identifying areas with high and low probability occurrences. The organization of the paper is as follows. Section 2 describes the statistical model and reviews the nonparametric estimator, as well as the bootstrap method used. In Section 3 the nonparametric spatial methods described in Section 2 are applied to two seismic data sets from two different regions. Section 4 presents the conclusions.

2 Nonparametric Regression Models

A regression model tries to describe the existing relation between an explicative variable, in general of dimension d, $\mathbf{X} = (X^1, X^2, ..., X^d)$, and a response variable

Y, based on a real data set $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$, obtained by the relationship:

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

where $m(\cdot)$ is the unknown regression function and ε_i are the random errors. Under the assumption of $\mathrm{E}(\varepsilon_i | \boldsymbol{X}_i) = 0$, if $\boldsymbol{X}$ and Y are random variables, the regression function at a location $\boldsymbol{x}$ is $m(\boldsymbol{x}) = E(Y | \boldsymbol{X} = \boldsymbol{x})$.

In this paper, the following spatial nonparametric regression model for the earthquake data will be used. Assume that a set of $\mathbb{R}^3$-valued random vectors, $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$, is observed in a specific time interval, where Y_i are scalar response variables and $\boldsymbol{X}_i$ are predictor variables with a common density f and compact support $\Omega \subseteq \mathbb{R}^2$. As stated in the Introduction, $\boldsymbol{X}_i$ will refer to the *epicenter locations* (latitude and longitude, expressed in degrees) corresponding to Y_i (magnitude).

The relationship between the locations and the response variable is assumed to be of the form (2), where $m(\cdot)$ is an unknown continuous and smooth function and ε is a second-order stationary process with:

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j | \boldsymbol{X}_i, \boldsymbol{X}_j) = C(\boldsymbol{X}_i - \boldsymbol{X}_j), \tag{3}$$

where $C(\mathbf{u})$ is a positive-definite function, called the covariogram (with $C(\mathbf{0}) = \mathrm{Var}(\varepsilon_i | \boldsymbol{X}_i) = \sigma^2$). The case of $C(\boldsymbol{u}) \equiv \sigma^2 \mathcal{I}_{\{\mathbf{0}\}}(\boldsymbol{u})$, where $\mathcal{I}_{\{\mathbf{0}\}}$ is the indicator function of the origin, corresponds to independent errors, whereas other covariance functions would allow for different types of spatial dependence. For instance, one of the best known covariogram families is the Matern class (e.g., Stein, 1999, pp.

48-52):

$$C_\theta(\boldsymbol{u}) = c_0 \mathcal{I}_{\{0\}}(\boldsymbol{u}) + \frac{c_1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\|\boldsymbol{u}\|}{a} \right)^{\nu} \mathcal{K}_\nu \left(\frac{\|\boldsymbol{u}\|}{a} \right), \tag{4}$$

where $\|\boldsymbol{X}\|$ is the Euclidean distance in $\mathbb{R}^d$, $\mathcal{K}$ is the modified Bessel function of the second type, c_0 is the nugget effect, c_1 is the partial sill (variance $\sigma^2 = c_0 + c_1$ 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. In geostatistics, covariogram models that allow a lack of continuity at the origin (non-zero nugget effect) are traditionally considered in practice, possibly corresponding to independent local variability (for more details see, Cressie, 1993, pp. 127-130).

This particular model assumes that the response variable Y is an unknown smooth function of location, "masked" by zero-mean (second-order stationary) errors. This is in contrast to the traditional Kriging models, which typically assume that the response variable is a simpler (linear or constant) function of location supplemented by spatially correlated noise. In these models, much of the observed behavior of the data is therefore attributed to noise instead of to the underlying mean function, and consequently, the error term should be modeled carefully (even considering a complicated spatial dependence structure). With a model like (2) assumptions about the error term are expected to have less effect on the conclusions obtained. Also, note that determining which model is in fact correct for the data cannot be done without replication at the same locations, and in practice, different

approaches could lead to similar estimated (or predicted) spatial maps. However, the interpretation of the map and the accompanying inference statements are different (for more details see, for instance, Altman, 1997, or Cressie, 1993, Section 3.1).

Obviously, we could consider a more general model if we include the earthquake depth in the variable $\boldsymbol{X}$, having a 3-dimensional explicative variable. Because one of the earthquake catalogues in our work (the Spanish data) does not always include the depth component for each recorded earthquake, and also given the difficulty of interpreting graphs in more than three dimensions, we have not included this variable in our analyses. Furthermore, model (2) can be modified to take into account additional spatio-temporal dimensions (including, for instance, the depth or even the temporal component) only through the errors ε_i, considering a more complicated spatio-temporal dependence (avoiding the so-called "curse of dimensionality" in nonparametric estimation).

Our first goal is to estimate the mean function $m(\cdot)$ using a nonparametric estimator. Nonparametric techniques try to estimate the regression function without assuming a specific parametric family of functions for $m(\cdot)$. Classic nonparametric regression estimators are based on explaining the relationship between the data using weighted local means, that is, the estimator of $m(\boldsymbol{x})$ can be written as:

$$\hat{m}_{\boldsymbol{H}}(\boldsymbol{x}) = \sum_{i=1}^{n} w_{\boldsymbol{H}}(\boldsymbol{X}, \boldsymbol{x}) Y_i,$$

where $\{w_{\boldsymbol{H}}(\cdot)\}_{i=1}^{n}$ is a weight sequence, characterized by a form and a scale. The

form corresponds to a *kernel function* $K : \mathbb{R}^d \to \mathbb{R}$, and the scale to a *smoothing parameter* or *bandwidth* $\boldsymbol{H}$.

While the election of the function K does not play an important role (normally, K is a density function with some regularity conditions) in the estimation (e.g., Härdle, 1990), bandwidth selection is more crucial, because the shape of the resulting estimator varies greatly depending on its value. In general, the bandwidth matrix $\boldsymbol{H}$ controls the shape and size of the local neighborhood used for estimating $m(\boldsymbol{x})$. So, if $\boldsymbol{H}$ is "small", we will obtain an undersmoothed estimation, with high variability; and, on the other hand, if $\boldsymbol{H}$ is "big", the resulting estimator will be very smooth and possibly with larger bias (see Section 2.2 below).

The importance of the bandwidth parameter in nonparametric estimation can be compared with that of the magnitude of completeness M_c in seismicity studies. The M_c magnitude estimation is directly correlated with the estimation of the b-value in the G-R law. As reported in Wiemer and Wyss (2002), "In the selection of M_c, we must choose a compromise that allows the coverage of the area of interest (or maximizes the area), but also does not unnecessarily reduce the number of events available". An underestimate of M_c may result in a too low b-value. If M_c is raised to large values, the uncertainty in the b-value estimate increases strongly. In this sense, an automatic way to calculate the M_c magnitude would be very helpful in seismic series research. We provide, in Section 2.2, an automatic method to choose the bandwidth for the data set at hand, in the sense that it avoids the problems above outlined due to the subjective selection of this parameter. This method has

been proven optimal, from a mathematically point of view (Francisco-Fernández and Opsomer, 2005).

2.1 Local Linear Regression for Spatial Data

When the explicative variables are bivariate, the local linear estimator for $m(\cdot)$ at a location $\boldsymbol{x}$ is the solution for γ to the least squares minimization problem

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

where $\boldsymbol{H}$ is a 2×2 symmetric positive definite matrix; K is a bivariate kernel and $K_{\boldsymbol{H}}(\boldsymbol{u}) = |\boldsymbol{H}|^{-1} K(\boldsymbol{H}^{-1}\boldsymbol{u})$. The Epanechnikov kernel function $K(\boldsymbol{x}) = \frac{2}{\pi} \max\left\{ \left(1 - \|\boldsymbol{x}\|^2\right), 0 \right\}$, a common choice for local linear regression, will be used throughout this article. The local linear regression estimator can be written explicitly as:

$$\hat{m}_{\boldsymbol{H}}(\boldsymbol{x}) = e_1^T \left(\boldsymbol{X}_{\boldsymbol{x}}^T \boldsymbol{W}_{\boldsymbol{x}} \boldsymbol{X}_{\boldsymbol{x}} \right)^{-1} \boldsymbol{X}_{\boldsymbol{x}}^T \boldsymbol{W}_{\boldsymbol{x}} Y \equiv s_{\boldsymbol{x}}^T Y, \tag{5}$$

where e_1 is a vector with 1 in the first entry and all other entries 0, $Y = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{W}_{\boldsymbol{x}} = \mathtt{diag}\left\{ K_{\boldsymbol{H}}(\boldsymbol{X}_1 - \boldsymbol{x}), ..., K_{\boldsymbol{H}}(\boldsymbol{X}_n - \boldsymbol{x} \right\}$, and

$$\boldsymbol{X}_{\boldsymbol{x}} = \begin{pmatrix} 1 & (\boldsymbol{X}_1 - \boldsymbol{x})^T \\ \vdots & \vdots \\ 1 & (\boldsymbol{X}_n - \boldsymbol{x})^T \end{pmatrix}.$$

Local linear regression has been broadly studied in the statistical context of univariate regression, and we refer to Wand and Jones (1995) for an overview. For bivariate local linear regression, Ruppert and Wand (1994) provide the relevant

asymptotic theory for the case in which the errors are independently distributed. In the spatial context, this assumption of independence is often not appropriate, and accounting for possible correlation is required for both inference and smoothing parameter selection. For a review on nonparametric regression with correlated errors see Hart (1996) and Opsomer *et al.* (2001). Recently, Francisco-Fernández and Opsomer (2005) discussed spatial smoothing and proposed a bandwidth selection method that allows for the presence of correlated errors.

A description of the bandwidth selection methods used to analyze the earthquake data will be presented in the following Section.

2.2 Bandwidth Selection

As stated above, smoothing parameter selection in kernel regression is important to be able to obtain reliable estimators. The influence of bandwidth $\boldsymbol{H}$ in the estimator of $m(\boldsymbol{x})$ can be observed in Figure 1, where three estimators of the regression function using the local linear estimator with three different bandwidths are presented. The data used to plot Figure 1 will be described and studied in Section 3.1. They consist of the epicenters and the corresponding magnitude of 1643 earthquakes occurring in the northwest Iberian Peninsula from 25/November/1944 to 03/April/2008 (with magnitude greater than or equal to 2).

We will use model (2) and the local linear regression estimator to map the estimated mean magnitude in that area. Each row of Figure 1 shows, in the left map, the epicenters and the bandwidth used to estimate the regression function at coordi-

nates 42.5° N and 8° W, and, in the right map, the corresponding estimator in that area computed with this bandwidth. The bandwidth used in the first row of Figure 1 is the bandwidth $\boldsymbol{H}_{GCVce}$, obtained using the "bias-corrected and estimated" GCV criterion, described below. The bandwidths used in the second and third rows are $0.5 \cdot \boldsymbol{H}_{GCVce}$ and $2 \cdot \boldsymbol{H}_{GCVce}$, respectively. More details on the parameters used to compute the bandwidth $\boldsymbol{H}_{GCVce}$ are found in Section 3.1. Compared with the *optimal* selection (first row), when a *small* bandwidth is used (second row), the estimator shows a significant amount of variation that appears to be spurious. Note that the variability shown here is actually less than that actually obtained by the regression, since the plot trims several large spikes that go well beyond the values on the scale, both in the positive and negative direction (done for comparability with the map in the first row of Figure 1). The estimator obtained with a larger bandwidth (third row) is much smoother than the *optimal* selection. This could lead to estimators removing notable parts of the mean pattern. An important observation about the estimators shown in Figure 1 (especially in the *optimal* case, first row) is that the extreme low and high values occur at the boundary of the estimation region, where smoothing methods (including the traditional approach of b-value mapping) often result in unreliable estimates. Although the use of local linear weights with a compactly supported kernel greatly minimizes these edge effects, the results obtained in the central part of the study area should be taken into account more.

[FIGURE 1 OVER HERE]

There are several methods to select the smoothing parameter. Some of these methods are especially designed for correlated observations, which are quite usual in the spatial smoothing context. In the case of spatial independent observations, classic smoothing parameters criteria for nonparametric regression are, for example, of cross-validation type (Craven and Whaba, 1979). These methods consider selecting the bandwidth $\boldsymbol{H}$ that minimizes the GCV function:

$$GCV(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{m}_{\boldsymbol{H}}(\boldsymbol{X}_i)}{1-\frac{1}{n}\mathrm{tr}(\boldsymbol{S})}\right)^2, \tag{6}$$

with $\boldsymbol{S}$ the $n \times n$ matrix whose ith row is equal to $\boldsymbol{s}^T_{\boldsymbol{X}_i}$, the smoother vector for $\boldsymbol{x} = \boldsymbol{X}_i$, and $\mathrm{tr}(\boldsymbol{S})$ is the 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 $\boldsymbol{H}$ can be achieved using numerical algorithms as implemented in statistical software. However, this criterion should not be used directly for bandwidth selection in the presence of correlated errors, because its expectation is severely affected by the correlation (Liu, 2001). In this case, we propose to use the "bias-corrected" GCV criterion, proposed in Francisco-Fernández and Opsomer (2005), based on selecting the bandwidth $\boldsymbol{H}$ that minimizes the function

$$GCV_c(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{m}_{\boldsymbol{H}}(\boldsymbol{X}_i)}{1-\frac{1}{n}\mathrm{tr}\left(\boldsymbol{S}\boldsymbol{R}\right)}\right)^2, \tag{7}$$

with $\boldsymbol{R}$ the correlation matrix of the errors.

In practice, matrix $\boldsymbol{R}$ is unknown, so, (7) is not yet a practical bandwidth selection criterion. Following Francisco-Fernández and Opsomer (2005) we will assume a parametric form for the covariogram, from which the correlogram can be obtained

$\rho_\theta(\boldsymbol{u}) = C_\theta(\boldsymbol{u})/C_\theta(\mathbf{0})$, and then replace the unknown $\boldsymbol{R}(\boldsymbol{\theta})$ in (7) by an estimate $\boldsymbol{R}(\hat{\boldsymbol{\theta}})$. This method is called the "bias-corrected and estimated" GCV criterion.

The theoretical optimality properties of this last criterion were discussed in Francisco-Fernández and Opsomer (2005). The authors considered the isotropic exponential model (corresponding to (4) with $c_0 = 0$ and $\nu = 0.5$) for the correlation function, and the dependence parameters were estimated using a simple method-of-moments estimators specifically obtained for this model. A similar approach will be taken here, but traditional geostatistical methods will be used in the dependence modeling (to avoid bias in the dependence parameter estimation).

The estimation of the spatial dependence was done through the variogram, $\gamma(\boldsymbol{u}) = C(\mathbf{0}) - C(\boldsymbol{u})$ (see, for instance, Cressie, 1993, Section 2.4.1, for an explanation on why variogram estimation is preferred to covariogram estimation). The general algorithm is as follows:

1. Obtain a pilot bandwidth matrix $\boldsymbol{H} = \boldsymbol{\Lambda}_{pilot}$ (for instance, using the standard GCV selection criterion (6)).

2. Using local linear estimator (5), with bandwidth matrix $\boldsymbol{H}$, obtain the residuals:

$$\hat{\varepsilon}_i = Y_i - \hat{m}_{\boldsymbol{H}}(\mathbf{X}_i), \quad i = 1, 2, \ldots, n.$$

3. Compute the classical variogram pilot estimator from the residuals:

$$\hat{\gamma}(\boldsymbol{u}_k) = \frac{1}{2|N(\boldsymbol{u}_k)|} \sum_{N(\boldsymbol{u}_k)} (\widehat{\varepsilon}_i - \widehat{\varepsilon}_j)^2, \quad k = 1, \ldots, K.$$

where $N(\boldsymbol{u}) = \{(i,j) : \boldsymbol{X}_i - \boldsymbol{X}_j \in Tol(\boldsymbol{u})\}$, $Tol(\boldsymbol{u})$ is a tolerance region around $\boldsymbol{u}$ and $|N(\boldsymbol{u})|$ denotes the number of contributing pairs at lag $\boldsymbol{u}$.

4. Fit a valid variogram model $\gamma_{\hat{\theta}}$ using a Weighted Least Squares (WLS) criterion:

$$\hat{\theta} = \arg\min_{\theta} \sum_{k=1}^{K} \omega_k \left(\gamma_\theta(\boldsymbol{u}_k) - \hat{\gamma}(\boldsymbol{u}_k)\right)^2 .$$

5. Update the bandwidth matrix $\boldsymbol{H}$, using the "bias-corrected and estimated" GCV method (7).

6. Repeat steps 2-5 until convergence is obtained.

In practice, few iterations of previous algorithm are usually needed (and very similar results were observed considering just one iteration). In step 4, following the recommendations of Journel and Huijbregts (1978, p. 74), the fit of a valid model is usually done up to half the maximum possible lag and considering only pilot estimations with at least 30 contributing pairs (i.e., $|N(\boldsymbol{u})| \geq 30$) The weights ω_i are usually chosen following the idea proposed by Cressie (1985) (i.e., $\omega_i = |N(\boldsymbol{u})|/\gamma_\theta(\mathbf{u}_i)$, and proceeding iteratively). In the examples shown in this work, a Matern model (4) was considered, allowing for the possibility of geometrical anisotropy (i.e., correlation depending on direction as well as on the distance between observations; see, Cressie, 1993, pp. 62-64). Note that the case of $\hat{c}_1 = 0$ corresponds with (estimated) independence. Other goodness-of-fit criteria could be used in step 4, for instance, the approach in Francisco-Fernández and Opsomer (2005) or maximum likelihood estimation.

2.3 Parametric Bootstrap

In general, bootstrap (Efron, 1979) is a resampling method that attempts to estimate the sampling distribution of a population by drawing new samples (with replacement) from the original data. Under certain probabilistic conditions (which are often true in reality), bootstrap usually produces more accurate and reliable results than traditional methods (see, for instance, Efron and Tibshirani (1993) for bootstrap and other resampling procedures). Nowadays, as a computer-aided statistical technique, bootstrap is largely used in many fields of applied statistics such as medicine, biostatistics, or seismology (Lamarre *et al.*, 1992; Pisarenko and Sornette, 2004).

In our case, we use the bootstrap to estimate the probability that an earthquake of magnitude above a fixed level occurs in a certain area. Bootstrap methods have been previously used in seismic hazard evaluation (Wiemer, 2001; Orlecka-Sikora, 2008), by approximating recurrence times, usually calculated through a parametric exponential expression with b-value estimates.

Now, to incorporate variability assessments in our analysis of earthquake magnitude, we extended the parametric *bootstrap* for correlated data discussed in Vilar-Fernández and González-Manteiga (1996). It follows the same steps:

1. Obtain, using the algorithm described in the previous Section, the optimal bandwidth matrix $\boldsymbol{H}$ (with the "bias-corrected and estimated" GCV criterion), the corresponding residuals $\hat{\varepsilon}_i$ and the covariance function parameter

estimates $\hat{\boldsymbol{\theta}}$.

2. Bootstrap data sets are generated by taking the estimated spatial trend $\hat{m}_{\boldsymbol{H}}(\boldsymbol{X}_i)$ and adding bootstrap errors generated as a spatially correlated set of errors. Bootstrap errors are obtained by the following steps:

   (a) Using $\hat{\boldsymbol{\theta}}$, compute the variance-covariance matrices of the residuals $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n)$, denoted by $\hat{\boldsymbol{V}}$.

   (b) Using Cholesky decomposition, find the matrix $\boldsymbol{P}$ such that $\hat{\boldsymbol{V}} = \boldsymbol{P}\boldsymbol{P}^T$.

   (c) Obtain the "independent" variables $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)$, given by

   $$\boldsymbol{e} = \boldsymbol{P}^{-1}\hat{\boldsymbol{\varepsilon}}.$$

   (d) These independent variables are centered and, from them, we obtain an independent bootstrap sample of sample size n, denoted by $\boldsymbol{e}^* = (e_1^*, e_2^*, \ldots, e_n^*)$.

   (e) Finally, the bootstrap errors $\hat{\boldsymbol{\varepsilon}}^* = (\hat{\varepsilon}_1^*, \ldots, \hat{\varepsilon}_n^*)$ are

   $$\hat{\boldsymbol{\varepsilon}}^* = \boldsymbol{P}\boldsymbol{e}^*$$

   and the bootstrap samples are

   $$Y_i^* = \hat{m}_{\boldsymbol{H}}(\boldsymbol{X}_i) + \hat{\varepsilon}_i^*,\ i = 1, 2, \ldots, n.$$

3. Once the bootstrap data sets are obtained, the above nonparametric regressions are repeated for each bootstrap sample using the same bandwidth $\boldsymbol{H}$

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). Finally, a map showing the frequency (across bootstrap replicates and for each location) of how often that location is included in the at-risk area is computed.

Note that this procedure is especially designed for when the errors are supposed to be spatially correlated. In the case of having independent errors, the previous algorithm can be easily simplified, since once the residuals $(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)$ are obtained (using in this case the GCV method for bandwidth selection) in the first step, the bootstrap errors $(\hat{\varepsilon}_1^*, \ldots, \hat{\varepsilon}_n^*)$ are directly obtained by resampling with replacement among the original residuals.

3 Data Analysis

3.1 Northwest Iberian Peninsula

The region studied in this Section is in the northwest part of the Iberian Peninsula. The statistical analysis is focused on data from the earthquakes occurring in the Asturian-Leonese and Galician-Castilian areas. These, mainly of Mesozoic and Tertiary age, are on the northwestern border of the Hesperic Massif (a continuous outcrop occupying most of the western half of the Iberian Peninsula. See, e.g., Julivert *et al.*, 1973; Julivert and Arboleya, 1984). The seismo-tectonics of the region

of interest is done in Rueda and Mezcua (2001), Martínez-Díaz *et al.* (2006) and Díaz *et al.* (2008). They divide the area into two families of faults showing alpine tectonic activity: Reverse ENE-WSW faults and Sinistral NNE-SSW faults. The whole area is limited by the coordinates 41° N – 44° N and 6° W – 10° W, involving the autonomic region of Galicia (Spain) and northern Portugal.

For the magnitude spatial analysis, we selected the data bank of the National Geographic Institute (IGN) of Spain, until April 2008 (see, Data and Resources Section). This public agency registers the seismic data in the Iberian Peninsula, Balearic and Canary Islands. Historically, in the NW region of Spain, only isolated events were reported (although on occasions with high intensity, including a 5.2 magnitude event in 1961) until the 1990's. Since then, the seismic activity increased considerably in Galicia, generating an important social alarm in the population, not accustomed to feeling earthquakes of a notable magnitude. Specifically, two seismic sequences of relative importance occurred in 1995 and 1997. These episodes have been widely analyzed in Rueda and Mezcua (2001) and Martínez-Díaz *et al.* (2006). In addition, the paper of Moreira (1991) describes the occurrence of tsunamis affecting the western Iberian coast, generated by earthquakes located around the Galician bank. The occurrence of an event of magnitude 4.5, 150 km away from the Galician coast, in 2005, stresses the results of this last work. Fault mechanisms have been analyzed in the mentioned works and others, such as Herraiz *et al.* (2000) or López-Fernández *et al.* (2004), related with the 1995 and 1997 episodes.

Figure 2 shows a map of this region, together with the selected data set, consist-

ing of $N = 1643$ earthquakes occurring from 25/November/1944 to 03/April/2008. Only earthquakes with magnitude above or equal to 2 were considered. We chose this value as the magnitude of completeness for this catalogue taking into account that several earthquakes with magnitude lower than two are dated before 1994. Until this year, the number of seismological stations in this area of Spain was very low, hence, the recordings have low reliability. Magnitude of completeness was computed by applying the EMR method (Woesner and Wiemer, 2005), using ZMAP software (Wiemer, 2001), with $M_c = 1.8$. We also computed its uncertainty by a bootstrap method, obtaining a value of 0.2.

As seen in Figure 2, small spatial groups are located in the areas described below. A study of the distribution of the inter-event times, as well as of temporary cluster grouping, can be seen in Estévez *et al.* (2002), for earthquakes occurring in 1987-2000.

[FIGURE 2 OVER HERE]

The fit of the nonparametric regression estimator (5) to the data set was done using the algorithm described in Section 2.2. To construct a map for the survey area of interest, the local linear estimates were computed on a regular 50×50 grid overlaying the field. As noted earlier, the adjusted GCV method for automatic bandwidth selection of Francisco-Fernández and Opsomer (2005) requires the specification of a model for the correlation. Following an exploratory analysis of the data, an isotropic exponential covariogram model (eq. (4) with $\nu = 0.5$) is apparently ad-

equate to describe the spatial dependence of the residuals. As already noted at the end of Section 2, this model specification is used in the selection of bandwidth values and does not determine the actual shape of the spatial distribution function $m(\boldsymbol{x})$. Therefore, modest differences between the true spatial correlation and the assumed correlation model would have a negligible effect on $\hat{m}(\boldsymbol{x})$. Figure 3 shows the pilot variogram estimations and the fitted variogram model. The estimated covariogram parameters were $\hat{c}_0 = 0.066$, $\hat{c}_1 = 0.132$ and $\hat{a} = 0.047$ (corresponding to a practical range of 0.141).

[FIGURE 3 OVER HERE]

Using these estimated parameters, the bandwidth matrix obtained with the selection criterion (7) was:

$$\boldsymbol{H} = \begin{bmatrix} 0.58 & 0 \\ 0 & 1.13 \end{bmatrix},$$

where the shape of this bandwidth at coordinates 42.5º N and 8º W, as well as the corresponding regression estimator in the area of interest, were previously shown in the first row of Figure 1. This bandwidth corresponds to a moderate amount of smoothing, since it implies that for any location $\boldsymbol{x}$ in the study region, 20-35% of the observations are contributing (have non-zero weight) to the nonparametric regression fit.

As observed in Figure 1 (first row), seismicity appears abundant and sparse, although some notable areas can be identified. Firstly, that defined at the Becerreá

zone (the area limited, approximately, by the coordinates 42.75º N – 43.5º N and 6.5º W – 8º W), where the two main episodes of 1995 and 1997 took place. This large number and size of earthquakes is attributed to a weak area on the SW-NE oriented Becerreá fault (Rueda and Mezcua, 2001; Martínez-Díaz *et al.*, 2002, 2006).

Some clusters of events can be observed around the 42th parallel (border between Spain and Portugal), without being too large in size. This area has a fault population mainly oriented N10º E–N60º E. The central-western part of Galicia shows moderate seismicity. The geological specific configuration of these last two areas can be seen in Ayarza *et al.* (1998). Finally, the Galicia Margin (offshore) is characterized by a diffuse seismicity band (Díaz *et al.*, 2008).

Now, we use the bootstrap method described in Section 2.3 to plot the pattern of probability of earthquake occurrences in the whole NW region. Figure 4 shows the maps with pointwise bootstrap probabilities of an earthquake (with magnitude above or equal to the considered threshold) occurring.

[FIGURE 4 OVER HERE]

To evaluate the sensitivity to the choice of the threshold, we considered two values: 2.5 in the left picture and 2.75 in the right one. There is a big difference between the two maps. So, although a large proportion of the area has a high probability of an earthquake of magnitude 2.5 or above occurring, only the areas of Becerreá and northern Portugal have a high hazard of an earthquake of magnitude 2.75 or higher occurring. Moreover, in this map, the highest values in the northern limit

are possibly due to a boundary effect, and these high values (close to one) possibly being spurious. As seen, our model takes into account the physical conditions along the area, which seems appropriate for forecasting seismicity.

One may think that the differences in Figures 4a and 4b are motivated by completeness reasons. Nevertheless, the main reason for this difference could be that the mean magnitude calculated with the optimal bandwidth over the whole region varies basically between 2 and 3, with no significant differences in the region. Hence, small changes in the threshold, from 2.5 to 2.75, can cause large differences in the plots of probabilities of the two Figures.

Finally, we checked the existence of variation in the mean magnitude between zones, but these were not statistically significant (as with Utsu's Test (Utsu, 1966) for b-values). More exactly, we could apply appropriated version of the Bowman test (Bowman, 2006) to check the differences between bivariate surfaces, but this is out of the scope of the present work.

3.2 California Data

With the aim of showing the behavior of our method in a completely different area to the one above, this methodology was applied to a seismic data set of California. Unlike the previous case, this zone has several differentiated regions and geological structures. The history of earthquakes in California is widely described in Toppozada and Branum (2004). Because, as they say: "The Bulletin of the Seismological Society of America has included many articles on various California earthquakes", we only

refer, e.g., to Petersen *et al.* (1996, 2008) for a detailed geographical description of this region and its seismic hazard.

For our example, we used the seismicity catalogue from the Advanced National Seismic System (ANSS) (see, Data and Resources Section). The selected grid is 30.08° N to 44.99° N in latitude and 125° W to 115° W in longitude. Earthquakes of magnitude above 3.5 for the period 01/January/1998 to 01/April/2008 were selected. The amount of data recorded during this period was $N = 4888$, considerably greater than that recorded in the Galician region previously analyzed. Figure 5 shows the study area and the earthquake epicenters.

[FIGURE 5 OVER HERE]

The magnitude of completeness for the selected period of time, using the EMR method, was 1.2. A graph of the frequency-magnitude distributions (obtained with the software ZMAP) is shown in Figure 6.

[FIGURE 6 OVER HERE]

In this case, we only considered earthquakes with magnitude above or equal to 3.5, mainly for computational time reasons. The number of necessary calculations in the nonparametric estimator (5) is linear in the sample size (n) and in the number of estimation points (a two-dimensional grid 50×50), but the cross-validation function (7) depends directly on the square of the number of observations. Also, the bootstrap procedure implies multiplying by B (number or bootstrap replications) the volume

of calculations. This kind of considerations was also used by Marsan and Lengliné (2008). Note that our method is not strongly affected by the election of magnitude of completeness. While accurate knowledge of M_c is essential for many seismicity-based studies, and particularly when mapping out seismicity parameters such as the b-value of the G-R relationship, in our procedure, it is not as important to obtain precise values of M_c.

For the California data set, the same steps as those used in the previous example were followed. First, we fit the local linear estimator to the earthquake magnitudes. Nevertheless, in this case, the spatial variability of the data was apparently captured by the trend, so the residuals seem to exhibit no spatial correlation. For example, Figure 7 shows an envelope for the empirical variogram obtained under spatial independence (by random permutation of the data values on the spatial locations). Almost all values lie inside the envelope, except the estimate at first lag. A closer look at the residuals showed that this was due to the presence of atypical observations (with respect to values in their neighborhood). Therefore, after the exploratory analysis of the data, the spatial independence of the data was considered reasonable.

[FIGURE 7 OVER HERE]

Due to this fact, we used the GCV algorithm for independent data, given in (6), to obtain a suitable bandwidth. The bandwidth obtained with this criterion was:

$$\boldsymbol{H} = \begin{bmatrix} 2.25 & 0 \\ 0 & 1.51 \end{bmatrix}.$$

So, with this bandwidth, estimates of the mean magnitude were computed in the area of interest. Then, the bootstrap method described in Section 2.3 was applied to obtain maps of estimates of the likelihood of an earthquake with magnitude above or equal to a threshold occurring at each location. Figure 8 shows the local linear regression estimator with that bandwidth, in the left panel, and the map with bootstrap probabilities, for threshold equal to 4.0, in the right panel.

[FIGURE 8 OVER HERE]

On the contrary to the area studied in the previous Section, the mean magnitude map shows a uniform pattern. This is due to a similar intensity rate of events (with magnitude equal to or greater than 4) over the whole rectangle. The bootstrap procedure helps us determine the areas with lower or higher forecasting probability (the last ones in black in the graph), as, for example, the one in eastern California, where the Hector Mine earthquake (16/10/1999, magnitude 7.1) took place. This agrees with Figure 8 (obtained for a different period of time) of Petersen et al. (1996).

Note also that when the nonparametric linear regression method is applied to a region like the one considered here, the results obtained could be thought to be in line with the gridded seismicity model, in which the earthquake rates determined

for cells are spatially smoothed using a Gaussian kernel, as done by Frankel (1995) and used by Hagos *et al.* (2006).

4 Conclusions

In this paper, we investigated a method for determining areas with high or low probability of occurrence of an earthquake of magnitude larger than or equal to a given threshold. The technique is a combination of nonparametric methods and a bootstrap algorithm. To estimate the mean magnitude in the area of interest, a nonparametric local linear estimator for bivariate observations was used. For this estimator, a matrix bandwidth is needed. In this paper, we suggest applying cross-validation methods for bandwidth selection from the data, that is, without subjective considerations. In spatial data analysis (that addressed in this paper), it is quite common to have spatial correlation between the observations. In that case, it is important to take this dependence into account in the bandwidth selection method. So, in a practical situation, before applying the method itself, it is necessary to check if the observations are or not spatially correlated. This can be done, for example, through semivariogram estimation. If the presence of spatial dependence is suspected, we propose the use of the "bias-corrected" GCV criterion. To apply this method, a parametric dependence structure for the data must be assumed and the corresponding parameters estimated from the observations. Once again, the estimated semivariograms can help with this process. The spatial depen-

dence structure is also needed in the resampling process in the bootstrap method. Then, if the observations can be assumed to be independent, the bandwidth matrix can be selected using the GCV method and the previous bootstrap method should be simplified for independent observations.

These methods were applied to two different earthquake data sets: the historic catalogue of the IGN for northwest Spain, and the earthquakes in California from January 1998 to April 2008. Different threshold values were used in each of these examples. The overall approach of spatial smoothing and bootstrap-based density mapping provides a useful and flexible set of statistical tools to simplify the visualization and analysis of the spatial distribution of earthquake data.

An important feature of our method is that we do not need to impose any assumption, either on the mechanism of the earthquake occurrence process or in its magnitude distribution, allowing the data "to speak for themselves". Firstly, our procedure to compute the mean magnitude using kernel estimation does not need of previous assumptions in the magnitude distribution. That is, our method remains valid if the magnitude distribution deviates from a standard exponential model, and the nonparametric estimation could be a good preliminary way to contrast the goodness of fit to this distribution. Moreover, our model considers the possibility of including spatial correlation between the data (through the errors). In the more typical way to visualize the spatial differences in the magnitude distribution by mapping the b-value, local fits using the k-nearest points to a grid point are carried out, k being a subjective value. This would be similar to using in our method a

local bandwidth (without considering the possibility of spatial correlation between the earthquakes) and assigning the same weights to the observations. This will possibly produce more variable estimations and cause a larger edge effect. Finally, while accurate knowledge of M_c is very important for mapping out the b-value of the G-R relationship, our procedure could be successfully applied even when this parameter is not accurately computed (now the bandwidth being the important parameter to select).

Secondly, the bootstrap method incorporates variability assessments of the magnitude estimates and allows us to approximate forecasting probabilities, without needing to suppose a prefixed form for the magnitude distribution (such as an exponential or pareto density) (Wiemer and Wyss, 1997; Kagan and Jackson, 2000).

The estimation approach seems helpful, because its application to two different areas (one with uniform moderate seismicity and other with high clustered activity) seems to correctly describe the features of the areas studied, and the procedure used is consistent with previous parametric and nonparametric studies.

Obviously, the depth in the spatial distribution should be also taken into account. The IGN catalogue (and even other catalogues) is very imprecise about this variable. Indeed, Díaz *et al.* (2008) suggested that perhaps Galicia was not well monitored by the permanent array of the National Network. Moreover, the difficulties indicated for plotting maps with the magnitude related with three variables, and also the extraordinary computational cost of the bootstrap procedure in this case, have motivated us to perform the nonparametric spatial distribution analysis

with only latitude and longitude as explicative variables.

Finally, our results remain in line with the universality of the seismic moment-frequency relation (Kagan, 1999). In that paper, Kagan showed that the universal constant (β-value; $\beta = 3/2b$) does not exhibit statistically significant regional variation in shallow seismic series. Similarly, the b-value (see Kagan formula 1) is generally around 1.0. However, variations from 0.5 to 1.5 have been reported, depending on several factors in a seismically active zone (Singh and Chadha, 2010 and references therein). In Kagan (2010) different arguments to find these differences can be seen, depending on the zone characteristics, the measurements or other factors.

5 Data and Resources

The data bank of the National Geographic Institute (IGN) of Spain can be searched using www.ign.es/ign/es/IGN/SisCatalogo.jsp.

The seismicity catalogue from the Advanced National Seismic System (ANSS) can be searched using http://quake.geo.berkeley.edu/anss/catalog-search.html.

The ZMAP (Wiemer, 2001) software was used to compute the completeness magnitudes in our data sets.

The MATLAB software, developed by The MathWorks (MathWorks, 2010), was used to implement the local linear regression estimator, the GCV and corrected GCV procedures as well as the bootstrap procedure used to generate the maps presented in the paper (code available upon request from the first author). Then, the plots in

Figures 3 and 7 as well as the WLS estimated covariogram parameters were obtained using the free statistical software R (R Development Core Team, 2010). Specifically, the geostatistical analysis was carried out using geoR package(Ribeiro and Diggle, 2001).

6 Acknowledgments

The research of Mario Francisco-Fernández has been partially supported by Grant MTM2008-00166 (ERDF included) and Grant PGIDIT07PXIB105259PR. The research of Alejandro Quintela-del-Río has been partially supported by MEC Grant MTM2006-03523. The authors thank the associate editor and two referees for constructive comments that improved the presentation of this article.

References

[1] Altman, N.S. (1997). Krige, smooth, both or neither? *ASA Proceedings of the Section on Statistics and the Environment*, 60–65, American Statistical Association, Alexandria, VA.

[2] Ayarza P., J.R. Martínez Catalán, J. Gallart, J.A. Pulgar, and J.J. Dañobeitia (1998). ESCIN 3.3: A seismic image of the Variscan crust in the interland of the NW Iberian massif, *Tectonics* **17** 171–186.

[3] Bowman, A.W. (2006). Comparing nonparametric surfaces, *Stat. Modelling* **6** 279–299.

[4] Choi, E., and P. Hall (1999). Nonparametric approach to analysis of space-time data on earthquake occurrences, *J. Comput. Graph. Stat.* **8** 733–748

[5] Craven, P., and G. Wahba (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numer. Math.* **31** 377–403.

[6] Cressie, N. (1985). Fitting variogram models by weighted least squares, *Math. Geol.* **17** 563–586.

[7] Cressie, N. (1993). *Statistics for Spatial Data.* John Wiley & Sons, New York, 2nd edition.

[8] Daley, D.J. and D. Vere-Jones (2003). *An Introduction to the Theory of Point Processes*, Springer-Verlag, New York.

[9] Díaz, J., J. Gallart, O. Gaspà, M. Ruiz, and D. Córdoba (2008), Seismicity analysis at the *Prestige* oil-tanker wreck area (Galicia Margin, NW of Iberia), *Mar. Geol.* **249** 150–165.

[10] Efron, B. (1979). Bootstrap methods: another look at the Jackknife, *Ann. Statist.* **7** 1–26.

[11] Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New-York.

[12] Estévez-Pérez, G., H. Lorenzo-Cimadevila, and A. Quintela-del-Río (2002). Nonparametric analysis of the time structure of seismicity in a geographic region, *Ann. Geophys.-Italy* **45** 497–512.

[13] Fan, J., and I. Gijbels (1996). *Local Polynomial Modeling and its Applications*, Chapman & Hall, London.

[14] Francisco-Fernández, M., and J.D. Opsomer (2005). Smoothing parameter selection methods for nonparametric regression with spatially correlated errors, *Canad. J. Statist.* **33** 539–558.

[15] Frankel, A. (1995). Mapping seismic hazard in the central and eastern United States, *Seismol. Res. Lett.* **66** 8–21.

[16] Grillenzoni, C. (2005). Non-parametric smoothing of spatio-temporal point processes, *J. Stat. Plan. Infer.* **128** 61–78.

[17] Gutenberg, B. and C.F. Richter, (1944), Frequency of earthquakes in California, *Bull. Seismol. Soc. Am.* **34** 185–188.

[18] Hagos, L., R. Arvidsson, and R. Roberts (2006). Application of the spatially smoothed seismicity and Monte Carlo methods to estimate the seismic hazard of Eritrea and the Surrounding region, *Nat. Hazards* **39** 395–418.

[19] Härdle, W. (1990). *Applied nonparametric regression,* Cambridge University Press.

[20] Hart, J. (1996). Some automated methods of smoothing time-dependent data, *J. Nonparametr. Stat.* **6** 115–142.

[21] Herraiz, M., G. De Vicente, R. Lindo-Ñaupari, J. Giner, J.L. Simón, J.M. González-Casado, O. Vadillo, M.A. Rodríguez-Pascua, J.I. Cicuéndez, A. Casas, L. Cabañas, P. Rincón, A.L. Cortés, M. Ramírez, and M. Lucini (2000). The recent (upper Miocene to Quaternary) and present tectonic stress distributions in the Iberian Peninsula, *Tectonics* **19** 762–786.

[22] Journel, A.G., and C.J. Huijbregts (1978). Mining geostatistics, Academic Press, London, 22–47.

[23] Julivert, M., J.M. Fontbote, A. Ribeiro, and L. Conde (1973). Mapa tectónico de la Península Ibérica y Baleares E: 1:1.000.000. *Inst. Geol. Min. Esp.*, Spain.

[24] Julivert, M., and M.L. Arboleya, (1984). A geometrical and kinematical approach to the nappe structure in an arcuate fold belt: the Cantabrian nappes (Hercynian chain, NW Spain), *J. Struct. Geol.* **6** 499–519.

[25] Kagan, Y.Y. (1990). Random stress and earthquake statistics: spatial dependence, *Geophys. J. Int.* **102** 573–583.

[26] Kagan, Y.Y. (1999). Universality of the seismic moment-frequency relation. *Pure Appl. Geophys.* **155** 537–573.

[27] Kagan, Y.Y. (2010). Earthquake size distribution: Power-law with exponent $\beta = 1/2$, *Tectonophysics* **490** 103–114

[28] Kagan, Y.Y., and D.D. Jackson (2000). Probabilistic forecasting of earthquakes, *Geophys. J. Int.* **143** 438–453.

[29] Kijko, A. (2004). Estimation of the maximum earthquake magnitude, $m_{\max}$, *Pure Appl. Geophys.* **161** 1655–1681.

[30] Kijko, A., and G. Graham (1998). Parametric-history procedure for probabilistic seismic hazard analysis. Part I: Estimation of Maximum Regional Magnitude $m_{\max}$, *Pure Appl. Geophys.* **152** 413–442.

[31] Lamarre, M., B. Townshend, and H.C. Shah (1992). Application of the bootstrap method to quantify uncertainty in seismic hazard estimates, *Bull. Seismol. Soc. Am.* **82** 104–119.

[32] Lasocki, S., and B. Orlecka-Sikora (2008). Seismic hazard assessment under complex source size distribution of mining-induced seismicity, *Tectonophysics* **456** 28–37.

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

[34] López-Fernández, C., J.A. Pulgar, J. Gallart, J.M. González-Cortina, J. Díaz, and M. Ruiz (2004). Sismicidad y tectónica en el área de Becerreá–Triacastela (Lugo, NO España), *Geogaceta* **36**, 51–54.

[35] Marsan, D., and O. Lengliné, (2008). Extending earthquakes' reach through cascading, Science **319** 1076–1079.

[36] Martínez-Díaz, J.J., R. Capote, M. Tsige, F. Martín-González, P. Villamor, and J.M. Insua (2002). Interpretación sismotectónica de las series sísmicas de Lugo (1995 y 1997): un caso de triggering en una zona continental estable, *Rev. Soc. Geol. España* **15** 201–215

[37] Martínez-Díaz, J.J., R. Capote, M. Tsige, P. Villamor, F. Martin-Gonzalez, and J.M. Insua-Arevalo (2006). Seismic triggering in a stable continental area: the Lugo 1995–1997 seismic sequences (NW Spain), *J. Geodyn.* **41** 440–449.

[38] Mathworks (2010). The MathWorks – MATLAB and Simulink for Technical Computing. www.mathworks.com.

[39] Moreira, V.S. (1991). Historical seismicity and seismotectonics of the area situated between the Iberian peninsula, Morocco, Selvagens and Azores Islands. In: Mezcua, J., Udias, A. (Eds.), Seismicity, Seismotectonics and Seismic Risk of the Ibero-Maghrebian Region, Publ. I.G.N., 8. Instituto Geográfico Nacional, Madrid, 213–225.

[40] Nadaraya, E. A. (1964). On Estimating Regression, *Theory Probab. Appl.* **9** 141–142.

[41] Ogata, Y. (1988). Statistical models for earthquake occurrence and residual analysis for point processes., *J. Amer. Statist. Assoc.* **83** 9–27.

[42] Ogata, Y. (1998). Space-time point-process models for earthquake occurrences, *Ann. Inst. Statist. Math.* **50** 379–402.

[43] Ogata, Y., and K. Katsura (1993). Analysis of temporal and spatial heterogeneity of magnitude frequency distribution inferred from earthquake catalogues, *Geophys. J. Int.* **113** 727–738.

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

[45] Orlecka-Sikora, B. (2008). Resampling methods for evaluating the uncertainty of the nonparametric magnitude distribution estimation in the Probabilistic Seismic Hazard Analysis, *Tectonophysics* **456** 38–51

[46] Parzen, E. (1962). On estimation of a probability density function and mode, *Ann. Math. Statis.* **32** 1065–1076.

[47] Petersen, M.D., W.A. Bryant, C.H. Cramer, T. Cao, N.S. Reichle, A.D. Frankel, J.J. Lienkaemper, P.A. McCrory, and D.P. Schwartz (1996). Probabilistic Seismic Hazard Assessment for the State of California, *USGS Open File Report* **706**.

[48] Petersen, M.D. A.D. Frankel, S.C. Harmsen, C.S. Mueller, K.M. Haller, R.L. Wheeler, R.L. Wesson, Y. Zeng, O.S. Boyd, D.M. Perkins, N. Luco, E.H. Field, C.J. Wills, and K.S. Rukstales (2008). Documentation for the 2008 Update of

the United States National Seismic Hazard Maps, *USGS Open File Report* **1128**.

[49] Pisarenko, V.F., and D. Sornette, (2004). Statistical detection and characterization of a deviation from the Gutenberg-Richter distribution above magnitude 8, *Pure Appl. Geophys.* **161** 839–864.

[50] R Developement Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

[51] Ribeiro Jr., P.J., and P.J. Diggle (2001). geoR: A package for geostatistical analysis. R-NEWS, **1** 15–18.

[52] Rueda, J., and J. Mezcua (2001). Sismicidad, sismotectónica y peligrosidad sísmica en Galicia, IGN Technical Publication 35.

[53] Ruppert, D., and M.P. Wand (1994). Multivariate locally weighted least squares regression, *Ann. Statist.* **22** 1346–1370.

[54] Simonoff, J. (1996). *Smoothing Methods in Statistics*, Springer-Verlag, New York.

[55] Singh, C., and R.K. Chadha (2010). Variations in the frequency–magnitude distribution of earthquakes with depth in the Koyna–Warna region, India, *J. Asian Earth Sci.* **39** 331–334.

[56] Stein, M. L. (1999). *Interpolation of spatial data, some theory for Kriging.* Springer Series in Statistics, Springer-Verlag, New York.

[57] Stock, C., and E. Smith (2002). Comparison between seismicity models generated by different kernel estimations. *Bull. Seismol. Soc. Am.* **92** 913–922.

[58] Toppozada, T. and D. Branum (2004). California earthquake history, *Ann. Geophys.* **47** 509–522.

[59] Torcal, F., A. Posadas, M. Chica, and I. Serrano (1999). Application of conditional geostatistical simulation to calculate the probability of occurrence of earthquakes belonging to a seismic series, *Geophys. J. Int.* **139** 703–725.

[60] Udias, A., and J. Rice (1975). Statistical analysis of microearthquake activity near San Andres Geophysical Observatory, Hollister, California, *Bull. Seismol. Soc. Am.* **65** 809–828.

[61] Utsu, T., (1966). A statistical significance test of the difference in b value between two earthquake groups, *J. Phys. Earth* **14** 37–40.

[62] Vere-Jones, D. (1978). Earthquake prediction - A statician's view, *J. Phys. Earth* **26** 129–146.

[63] Vere-Jones, D. (1992). Statistical methods for the description and display of earthquake catalogues, in *Statistics in the Environmental and Earth Sciences*, Walden, A.T. and P. Guttorp (Editors), London 220–244.

[64] Vilar-Fernández, J.M., and W. González-Manteiga (1996). Bootstrap test of goodness of fit to a linear model when errors are correlated, *Comm. Statist. Theory Methods* **25** 2925–2953.

[65] Wand, M. P., and M. C. Jones (1995). *Kernel Smoothing*, Chapman and Hall, London.

[66] Watson, G.S., and M.R. Leadbetter (1964). Hazard analysis I, *Biometrika* **51** 175–184.

[67] Wiemer, S. (2001). A software package to analyze seismicity: ZMAP, *Seismol. Res. Lett.* **72** 374–383.

[68] Wiemer, S., and J. Benoit (1996). Mapping the b-value anomaly at 100 km depth in the Alaska and New Zealand subduction zones, *Geophys. Res. Lett.* **23** 1557–1560.

[69] Wiemer, S., and M. Wyss (1997). Mapping the frequency-magnitude distribution in asperities: An improved technique to calculate recurrence times?, *J. Geophys. Res.* **102** 15115–15128.

[70] Wiemer, S., and M. Wyss (2002). Mapping spatial variability of the frequency-magnitude distribution of earthquakes, *Adv. Geophys.* **45**, 259–302.

[71] Woessner, J., and S. Wiemer (2005), Assessing the quality of earthquake catalogues: estimating the magnitude of completeness and its uncertainty, *Bull. Seismol. Soc. Am.* **95** 684–698.

[72] Zhuang, J., Y. Ogata, and D. Vere-Jones (2002). Stochastic declustering of space-time earthquake occurrences, *J. Am. Stat. Assoc.* **97** 369–380.

Authors' affiliations, Addresses

Mario Francisco-Fernández. Universidad de A Coruña. Departamento de Matemáticas, Facultad de Informática, Campus de Elviña s/n, A Coruña, 15071, Spain.
E-mail: *mariofr@udc.es*

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

Rubén Fernández Casal. Universidad de A Coruña. Departamento de Matemáticas, Facultad de Informática, Campus de Elviña s/n, A Coruña, 15071, Spain.
E-mail: *rfcasal@udc.es*

Figure Captions

Figure 1: Epicenter of earthquake locations, bandwidths and regression estimators computed with three different bandwidths.

Figure 2: Epicenters of earthquake locations used in the northwest Iberian Peninsula example.

Figure 3: Classical semivariogram estimations and WLS fitted exponential model.

Figure 4: Maps with bootstrap probabilities of areas with seismic hazard for different threshold values.

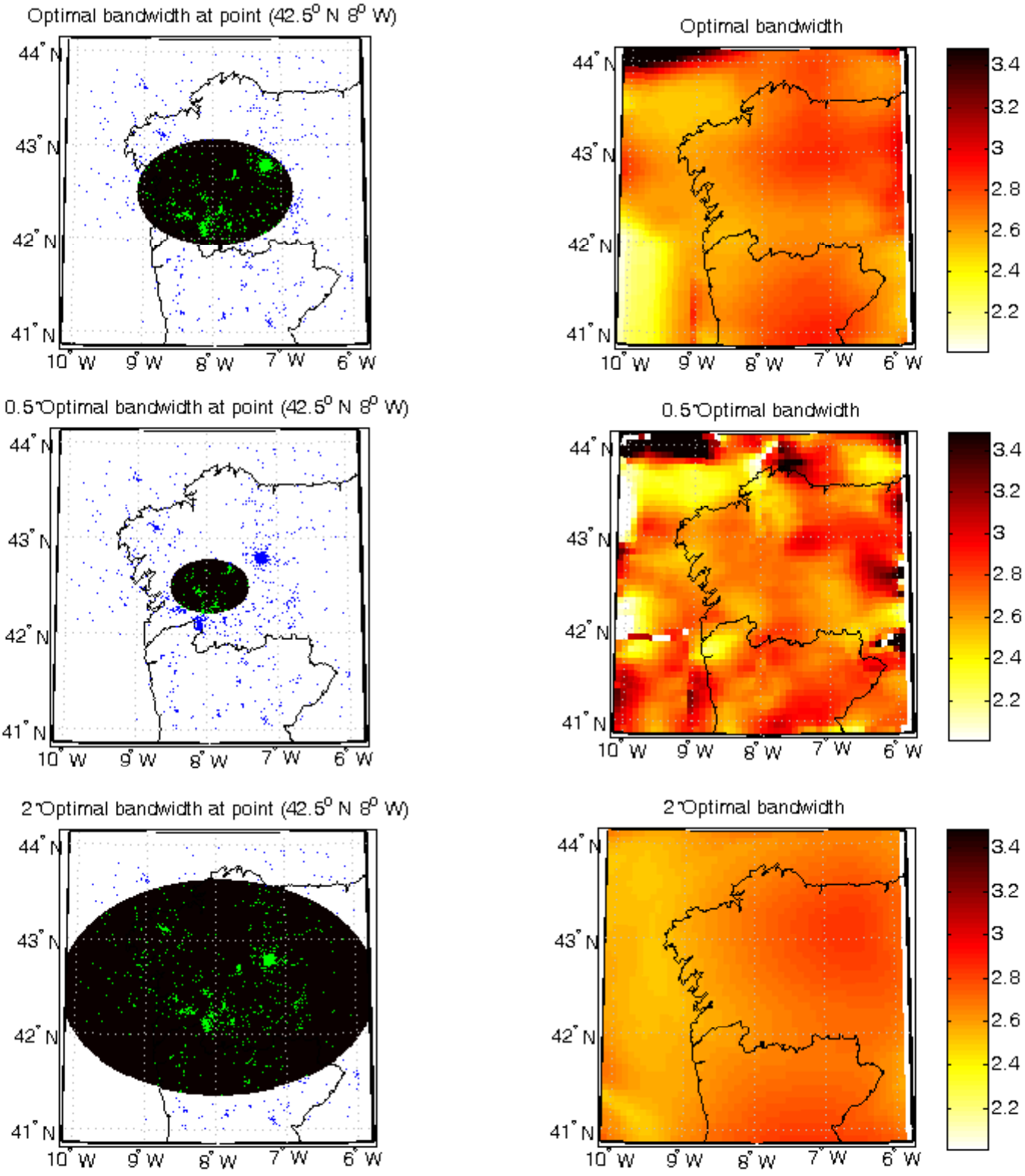
Figure 5: Epicenters of earthquakes used in the California example.

Figure 6: A magnitude completeness graph obtained using the ZMAP program in the California example.

Figure 7: Envelope for the empirical variogram obtained under spatial independence.

Figure 8: Estimated mean magnitude (left panel) and estimated probability of an earthquake of magnitude 4 or above occurring (right panel).

Figure 1


Optimal bandwidth at point (42.5° N 8° W)
Optimal bandwidth
0.5° Optimal bandwidth at point (42.5° N 8° W)
0.5° Optimal bandwidth
2° Optimal bandwidth at point (42.5° N 8° W)
2° Optimal bandwidth
44° N
43° N
42° N
41° N
10° W
9° W
8° W
7° W
6° W
3.4
3.2
3
2.8
2.6
2.4
2.2

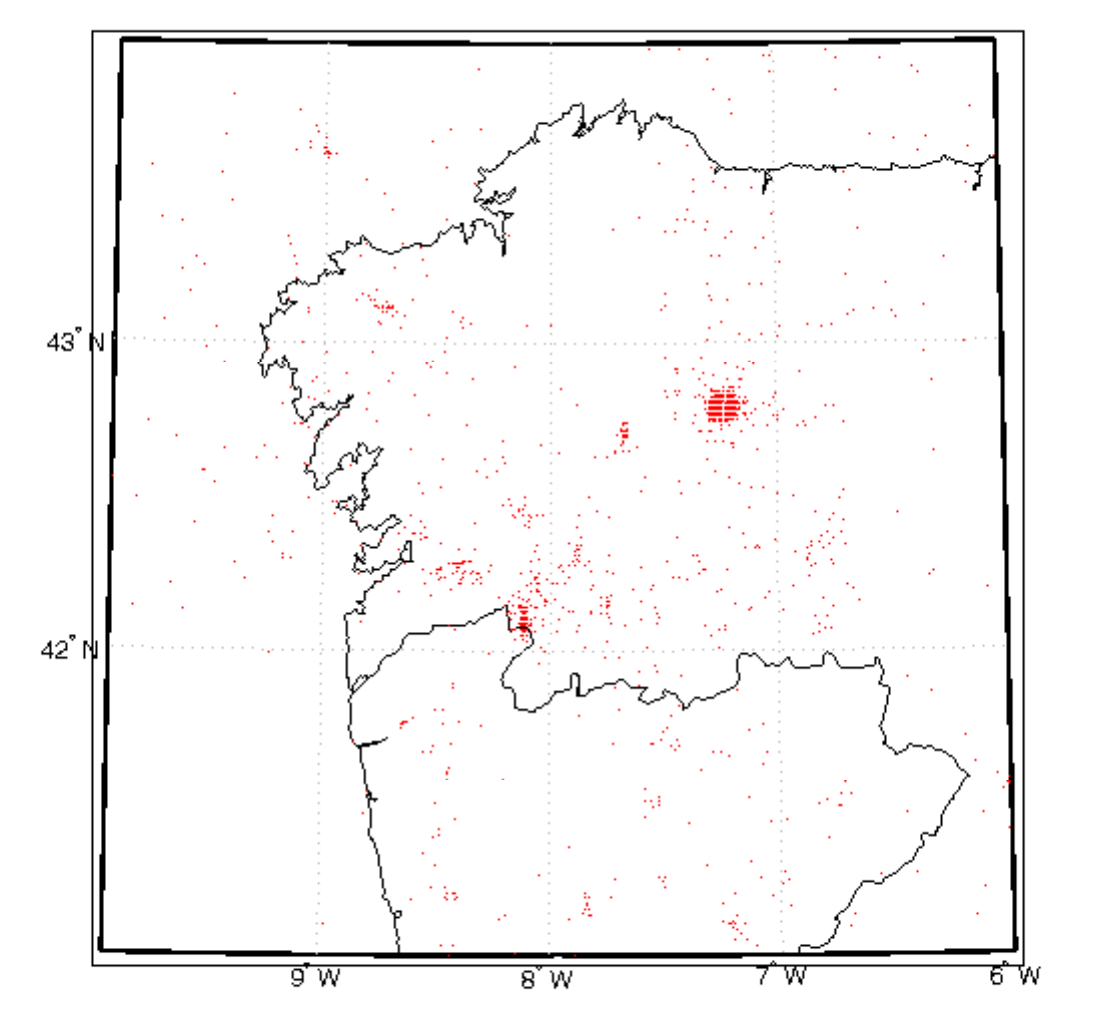
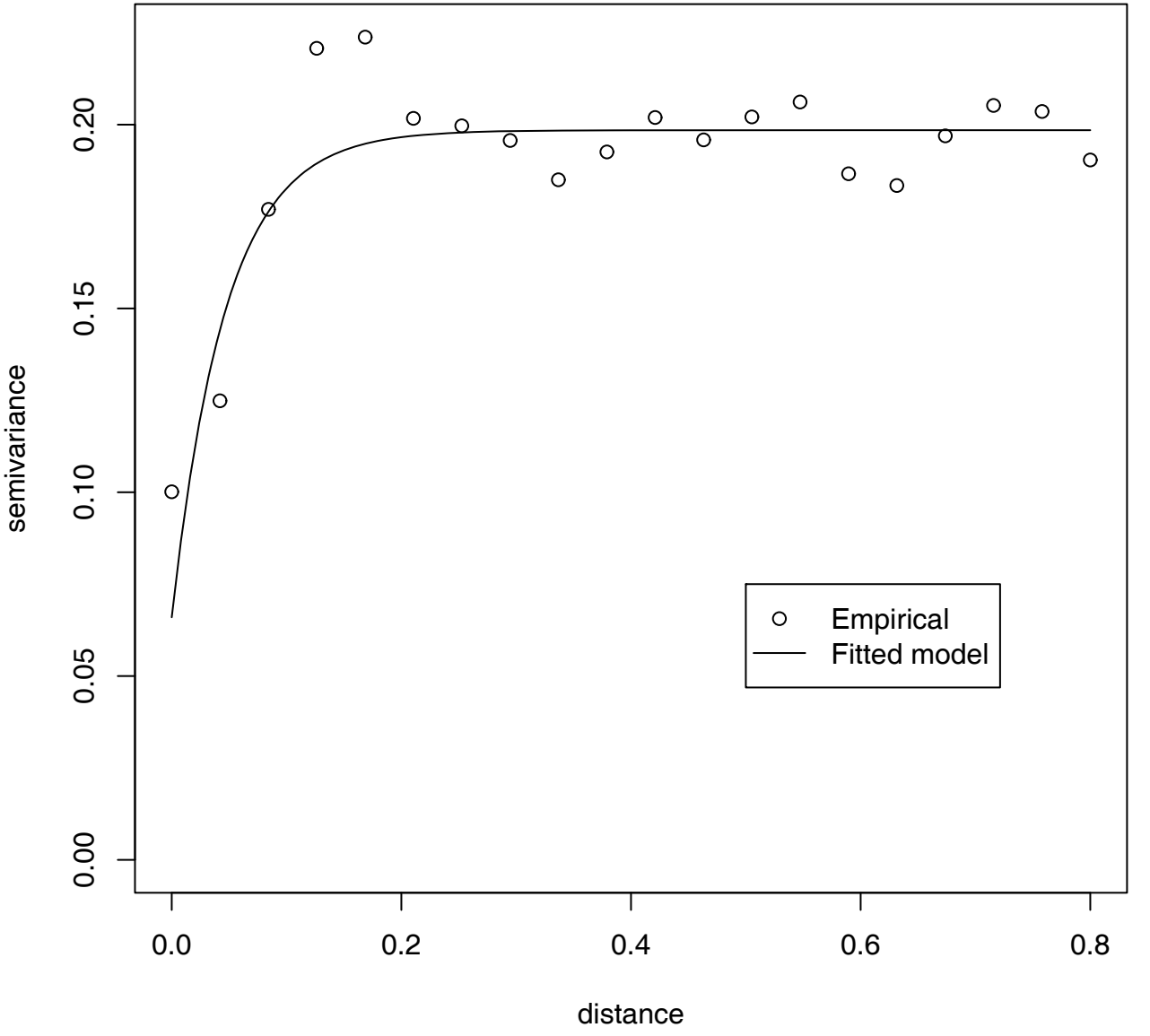
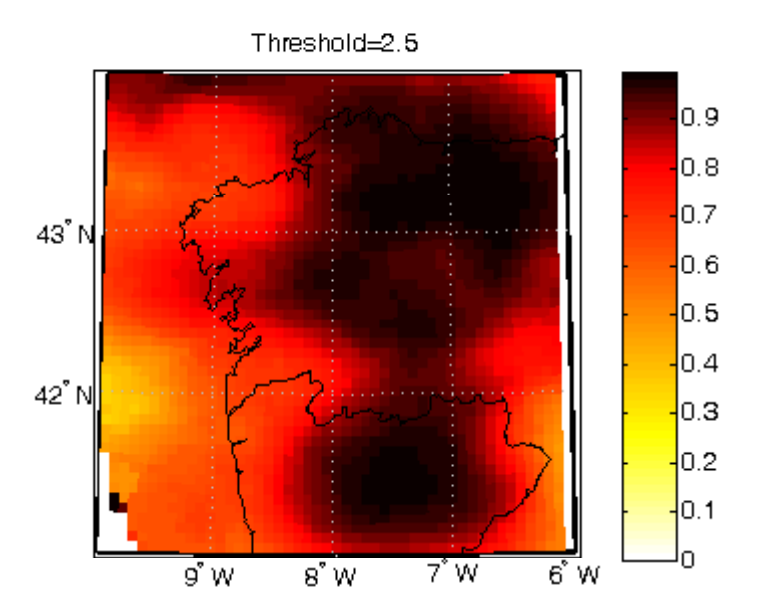
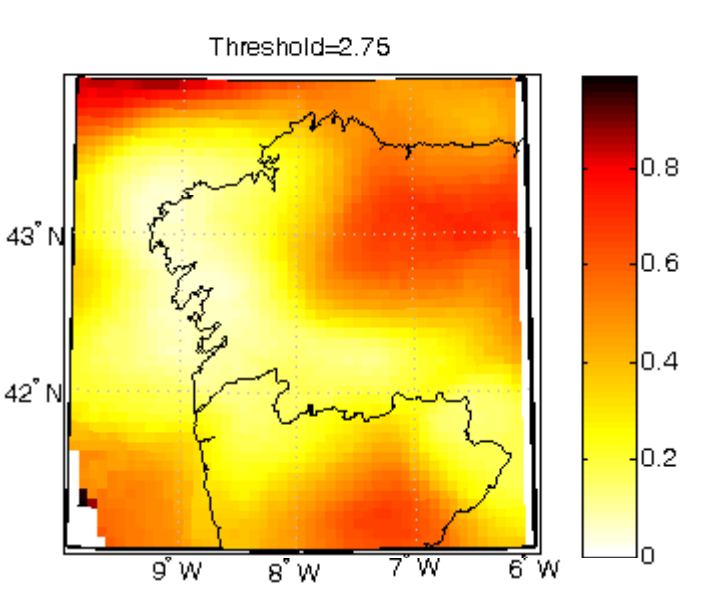
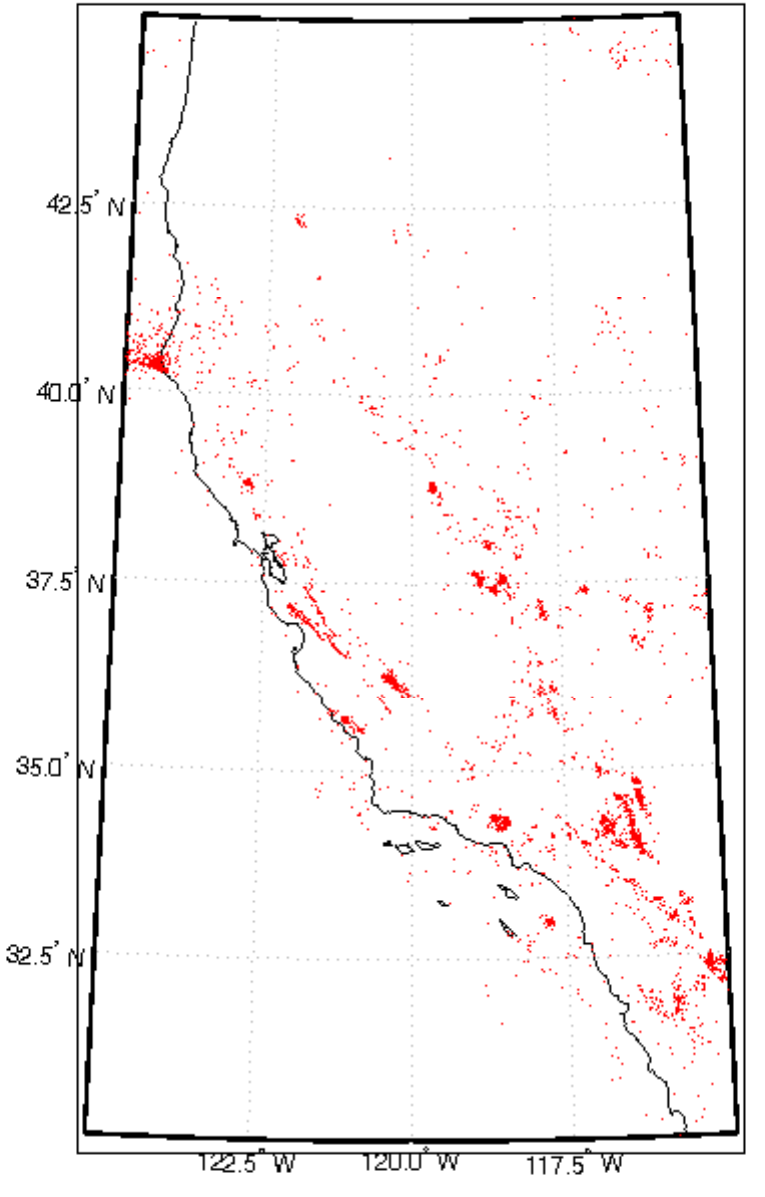

43° N
42° N
9° W
8° W
7° W
6° W

Figure 3


semivariance
0.00
0.05
0.10
0.15
0.20
0.0
0.2
0.4
0.6
0.8
distance
Empirical
Fitted model


Threshold=2.5
43° N
42° N
9° W
8° W
7° W
6° W
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0


Threshold=2.75
43° N
42° N
9° W
8° W
7° W
6° W
0.8
0.6
0.4
0.2
0


42.5° N
40.0° N
37.5° N
35.0° N
32.5° N
122.5° W
120.0° W
117.5° W

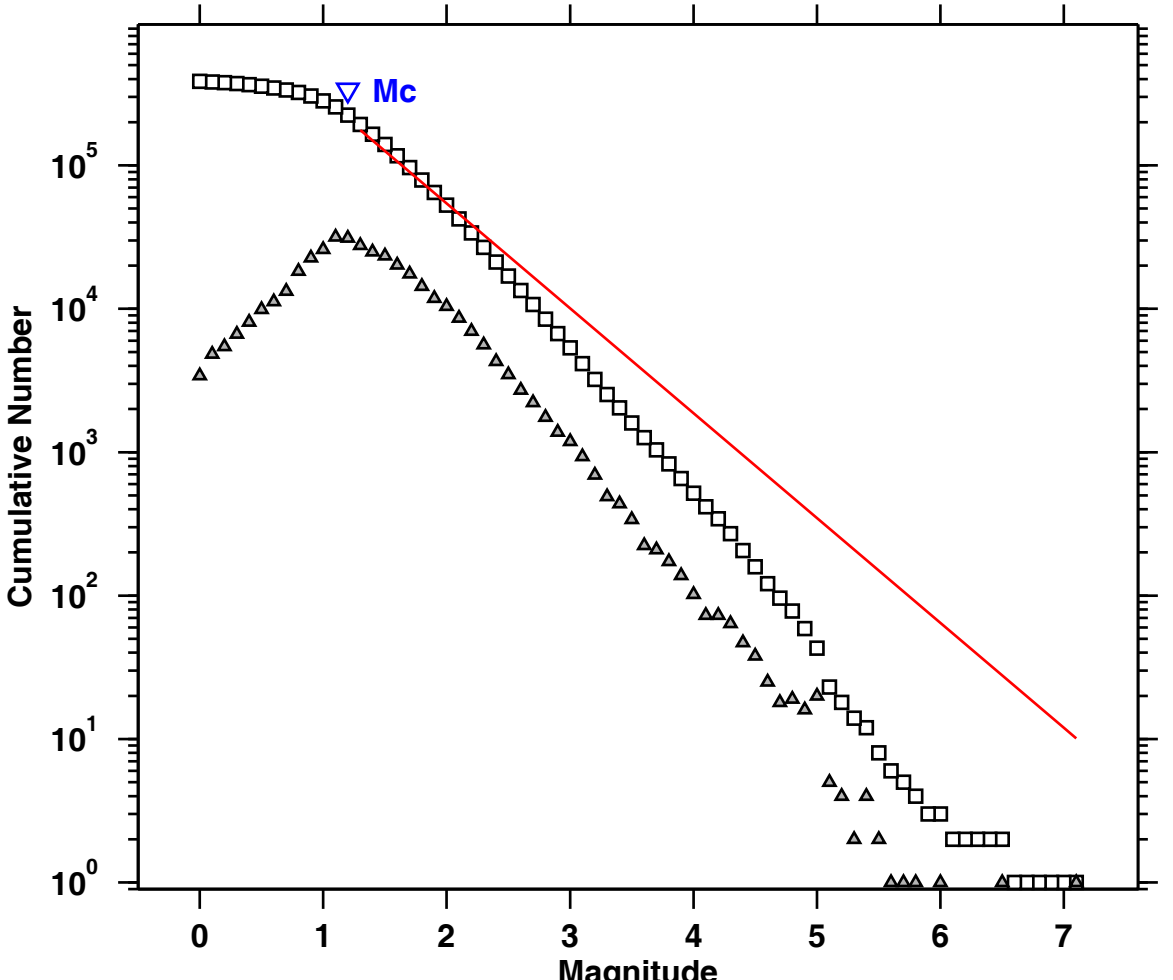
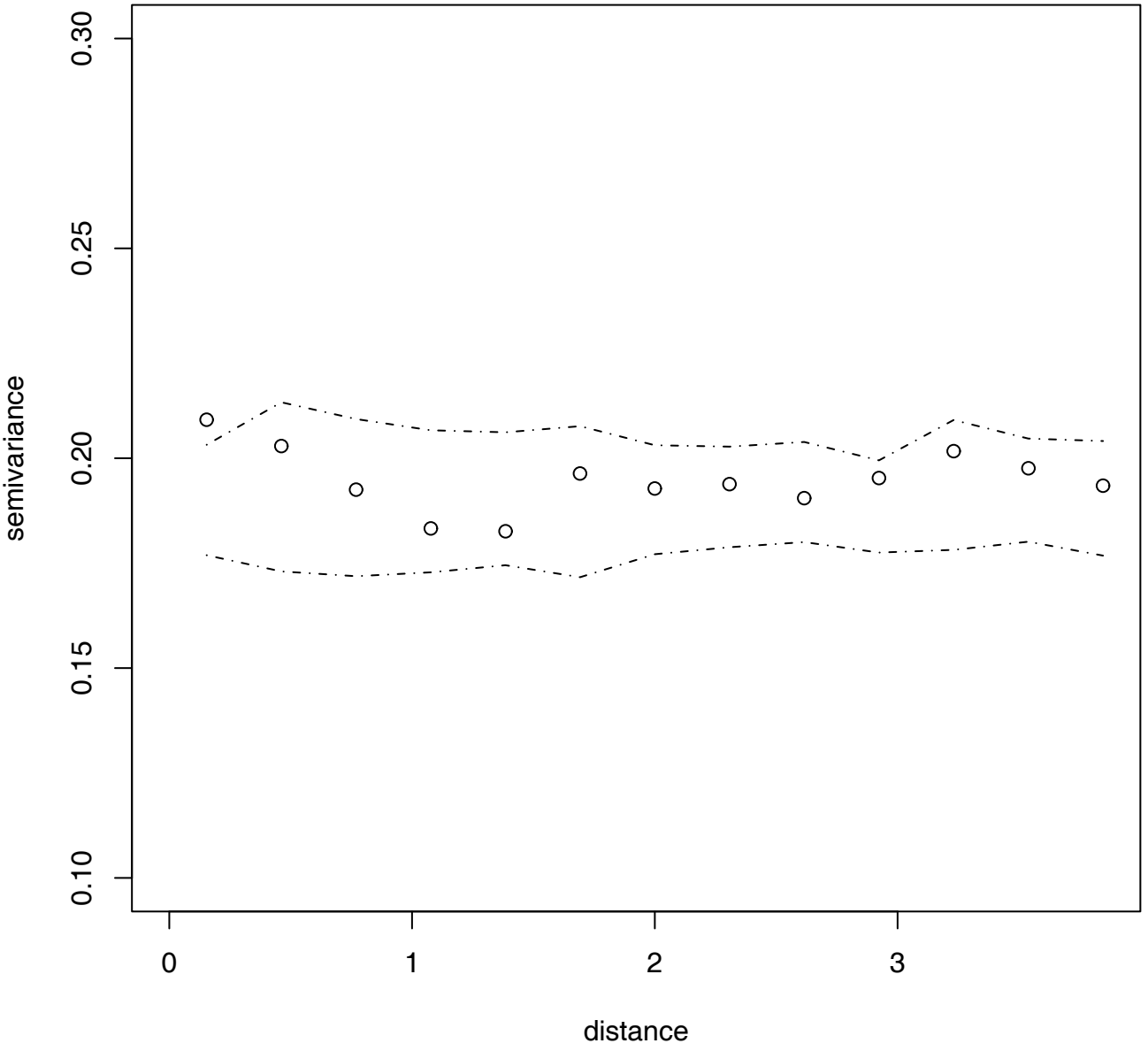
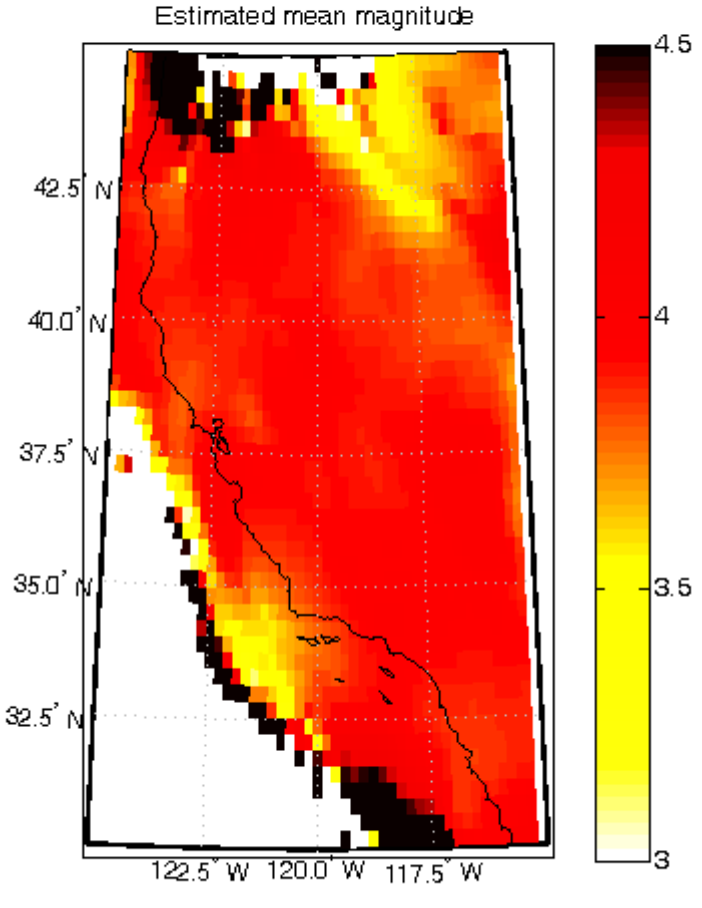
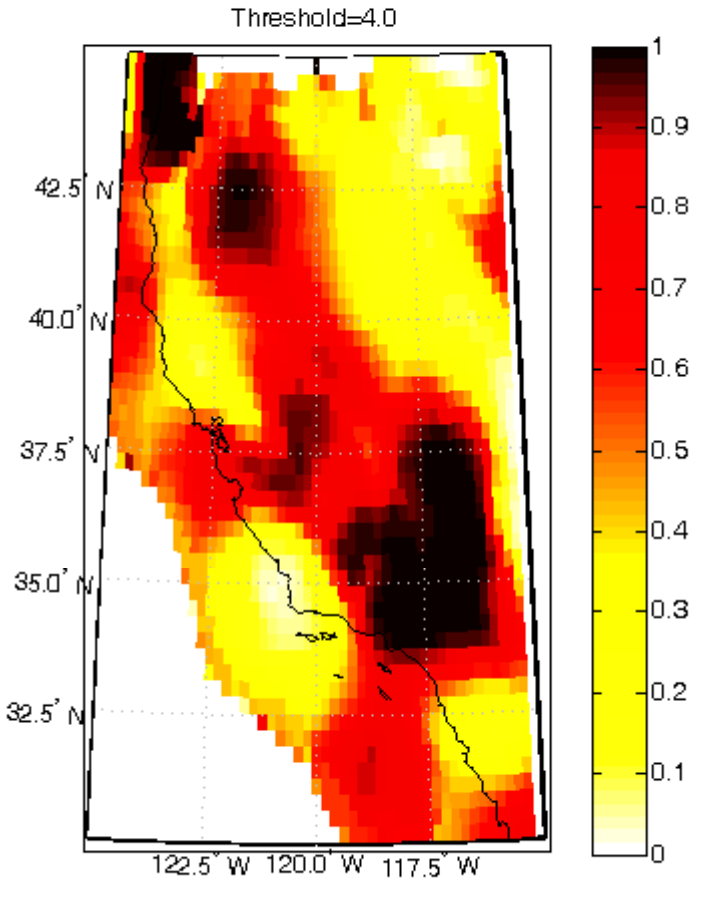

Mc
Cumulative Number
10^5
10^4
10^3
10^2
10^1
10^0
0
1
2
3
4
5
6
7
Magnitude

Figure 7




Estimated mean magnitude
42.5° N
40.0° N
37.5° N
35.0° N
32.5° N
122.5° W
120.0° W
117.5° W
4.5
4
3.5
3


Threshold=4.0
42.5° N
40.0° N
37.5° N
35.0° N
32.5° N
122.5° W
120.0° W
117.5° W
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0