
# Nonparametric local linear regression applied to the spatial structure of seismic events magnitude

M. Francisco-Fernández[1], A. Quintela-del-Río[2,*] and R. Fernández-Casal[3]

[1] *Departamento de Matemáticas, Universidad de A Coruña; mariofr@udc.es*
[2] *Departamento de Matemáticas, Universidad de A Coruña; aquintela@udc.es*
[3] *Departamento de Matemáticas, Universidad de A Coruña; rfcasal@udc.es*
[*] *Corresponding author*

***Abstract.*** *We describe the spatial structure of the earthquakes magnitude in a concrete geographical zone, by means of the nonparametric local polynomial regression estimator. We propose to use a bandwidth selection method 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 that shows the probability of being at high and low seismic risk in the considered area. These techniques are applied to an earthquakes data set of the Galician region (Spain).*

***Keywords.*** *Earthquakes; 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 as 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 of the epicenter, $X_i^3$ the depth of the focus, $t_i$ the origin time and $Y_i$ the magnitude. In this paper we suggest the application of nonparametric methods for analyzing seismic data. More concretely, we are interested in mapping the (bidimensional) spatial distribution of the earthquakes magnitudes, by means of the model

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

where $m(\cdot)$ is a regression function, in which we do not suppose any concrete parametric model and $\varepsilon_i$ are random errors that may or may not be spatially correlated. Here, we show the utility of a particular

type of nonparametric method, the named "local linear regression" [3], to the spatial statistical analysis of earthquakes data. We will do this by considering the problem of visualizing the spatial pattern of earthquakes magnitude, and applying to a seismic data set. The representation of the magnitude as a smooth spatial function, jointly with the use of bootstrap techniques, provides a useful "first step" in identifying areas of high and low seismic risk. The organization of this extended abstract is the following: Section 1 describes the statistical model, reviews the nonparametric estimator, the bandwidth selection method used, as well as the bootstrap method used. Section 2 provides information on the study area and data, and shows the behavior of the nonparametric spatial methods on those real earthquake data.

# 2 Statistical methods

## 2.1 Local linear regression for spatial data

In this work, the following spatial nonparametric regression model for the earthquakes data will be used. Assume that a set of $\mathbb{R}^3$-valued random vectors, $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$, are observed in a concrete interval of time, where the $Y_i$ are scalar responses variables and the $\boldsymbol{X}_i$ are predictor variables with a common density $f$ and compact support $\Omega \subseteq \mathbb{R}^2$. We will refer to the $\boldsymbol{X}_i$ as the *locations* (latitude and longitude, expressed in degrees) corresponding to the $Y_i$ (magnitude). The relationship between the locations and the responses variable is assumed to be of the form (1), where $m(\cdot)$ is an unknown continuous and smooth function, and $\varepsilon$ is a second order stationary process with:

$$\mathrm{Cov}\left(\varepsilon_i, \varepsilon_j \middle| \boldsymbol{X}_i, \boldsymbol{X}_j\right) = C(\boldsymbol{X}_i - \boldsymbol{X}_j), \tag{2}$$

where $C(\mathbf{u})$ is a positive-definite function, called the covariogram (with $C(\mathbf{0}) = \mathrm{Var}\left(\varepsilon_i \middle| \boldsymbol{X}_i\right) = \sigma^2$).

Our first goal is to estimate the mean function $m(\cdot)$ using a nonparametric estimator. Classical nonparametric regression estimators are based on explaining the relationship between the data by using weighted local means, that is, the estimator of $m(\boldsymbol{x})$ can be written as

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

When the explicative variables are bivariate variables, the local linear estimator for $m(\cdot)$ at a location $\boldsymbol{x}$ is the solution for $\gamma$ 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 $2 \times 2$ symmetric positive definite matrix; $K$ is a bivariate kernel and $K_{\boldsymbol{H}}(\boldsymbol{u}) = |\boldsymbol{H}|^{-1} K(\boldsymbol{H}^{-1}\boldsymbol{u})$.

## 2.2 Bandwidth selection

Here, we propose to use a modified version of the generalized cross-validation (GCV) criterion [1], called the "bias-corrected" GCV criterion proposed in [4], based on selecting the bandwidth $\boldsymbol{H}$ that minimizes the function

$$GCV_c(\boldsymbol{H}) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{Y_i - \hat{m}_{\boldsymbol{H}}(\boldsymbol{X}_i)}{1 - \frac{1}{n}\mathrm{tr}(\boldsymbol{S}\boldsymbol{R})} \right)^2, \tag{3}$$

with $\boldsymbol{R}$ the correlation matrix of the errors and $\boldsymbol{S}$ the $n \times n$ matrix whose $i$th row is equal to the smoother vector for $\boldsymbol{x} = \boldsymbol{X}_i$, and tr($\boldsymbol{SR}$) is the trace of matrix $\boldsymbol{SR}$.

In practice, matrix $\boldsymbol{R}$ is unknown, so that, (3) is not yet a practical bandwidth selection criterion. Following [4] we will assume a parametric form for the covariogram, from which the correlogram $\rho_\theta(\boldsymbol{u}) = C_\theta(\boldsymbol{u})/C_\theta(\boldsymbol{0})$ can be obtained, and then replace the unknown $\boldsymbol{R}(\boldsymbol{\theta})$ in (3) 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 [4]. A similar approach will be used here, but traditional geostatistical methods will be employed in the dependence modelling (in order to avoid bias in the dependence parameter estimation). The estimation of the spatial dependence was done through the variogram, $\gamma(\boldsymbol{u}) = C(\boldsymbol{0}) - C(\boldsymbol{u})$ (see e.g. [2], section 2.4.1, for a explanation about why variogram estimation is preferred to covariogram estimation). The description of the general algorithm can be seen in [5].

### 2.3 Parametric bootstrap

Now, in order to incorporate variability assessments in our analysis of the earthquakes magnitude, we extended the parametric *bootstrap* for correlated data discussed in [6]. It follows the steps detailed in [5], and is especially designed for when the errors are supposed to be spatially correlated.

## 3 Earthquakes analysis

Our area of interest focus on the Northwest of the Iberian Peninsula, concretely the area limited by the coordinates 42º N – 44º N and 6º W – 10º W, that involves the autonomic region of Galicia. We have selected the data bank of the National Geographic Institute (IGN) of Spain (until April of 2008) at http://www.ign.es/ign/es/IGN/SisCatalogo.jsp.

Following an exploratory analysis of the data, it could be seen that an isotropic exponential covariogram model is apparently adequate to describe the spatial dependence of the residuals. This model specification is used in the selection of bandwidth values and does not determine the actual shape of the spatial distribution function $m(\boldsymbol{x})$.

The bandwidth matrix obtained with the selection criterion (3) was:

$$\boldsymbol{H} = \begin{bmatrix} 0.54 & 0 \\ 0 & 1.13 \end{bmatrix},$$

This bandwidth corresponds to a moderate amount of smoothing, since this bandwidth matrix implies that for any location $\boldsymbol{x}$ not on the boundary of the study region, 20-40% of the observations are contributing (have non-zero weight) to the nonparametric regression fit. Next, we use the bootstrap method described before to plot maps of estimates of the likelihood of an earthquake with magnitude larger or equal than a threshold occurring in each location of the area of interest. Figure 1 show the maps with pointwise bootstrap probabilities of being considered at risk of occurring an earthquake with magnitude larger or equal than the threshold considered. To evaluate the sensitivity to the choice of the

threshold, we considered two thresholds: 2.5 in the left picture and 2.75 in the right picture. We can observe that there is an important difference between both maps. So, while a big proportion of the area is in high risk of occurring an earthquake with a magnitude larger or equal than 2.5, only the area limited, approximately, by the coordinates 42.75º N – 43.5º N and 6.5º W – 8º W has an important risk of occurring an earthquake with a magnitude larger or equal than 2.75. On the other hand, in this map, the highest values in the North limit are possibly due to a boundary effect, making it likely these very high risk values (close to one) are indeed spurious.

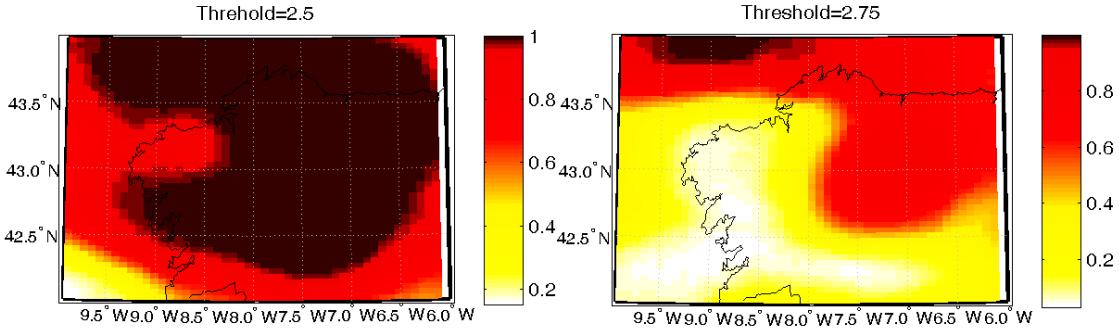


Figure 1: Maps with bootstrap probabilities of areas with seismic risk for different threshold values.

**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.

## References

[1] Craven, P. and Wahba, G. (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.

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

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

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

[5] Francisco-Fernández, M., Quintela-del-Río, A. and Fernández-Casal, R. (2010), A nonparametric analysis of the spatial distribution of earthquakes magnitude, *submitted.*

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