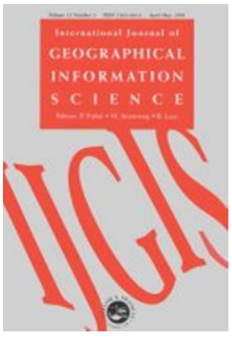

Testing spatial heterogeneity in geographically weighted principal component analysis

Journal:	*International Journal of Geographical Information Science*
Manuscript ID	IJGIS-2015-0602.R2
Manuscript Type:	Research Article
Keywords:	Principal components, Kernel smoothing, Bandwidth selection, Soil contamination

International Journal of Geographical Information Science
Vol. 00, No. 00, Month 200x, 1–20

RESEARCH ARTICLE

Testing spatial heterogeneity in geographically weighted principal components analysis

Javier Roca-Pardiñas[a], Celestino Ordóñez [b*], Tomás R. Cotos-Yáñez[a] , and Rubén Pérez-Álvarez[c]

[a]*Department of Statistics , and Operations Research, University of Vigo, 36208 Vigo,Spain*; [b]*Department of Mining Exploitation and Prospecting, University of Oviedo, 33600, Oviedo, Spain*; [c]*Department of Transports , and Technology of Projects and Processes, University of Cantabria, 39316 Torrelavega, Spain*

(*Received 00 Month 200x; final version received 00 Month 200x*)

We propose a method to evaluate the existence of spatial variability in the covariance structure in a geographically weighted principal components analysis (GWPCA). The method, that is extensive to locally weighted principal components analysis (LWPCA), is based on performing a statistical hypothesis test using the eigenvectors of the PCA scores covariance matrix. The application of the method to simulated data shows that it has a greater statistical power than the current statistical test that uses the eigenvalues of the raw data covariance matrix. Finally, the method was applied to a real problem whose objective is to find spatial distribution patterns in a set of soil pollutants. The results show the utility of GWPCA versus PCA.

Keywords: Principal components; Kernel smoothing; Bandwidth selection; Soil contamination

*Corresponding author. Email: ordonezcelestino@uniovi.es

ISSN: 1365-8816 print/ISSN 1362-3087 online

DOI: 10.1080/1365881YYxxxxxxxx
http://www.informaworld.com

1. Introduction

GWPCA, as well as LWPCA, are extensions of standard global PCA when the covariance structure of the data is not supposed to be constant through space (Tipping and Bishop 1999, Charlton *et al.* 2010, Harris *et al.* 2011). The idea is similar to that of geographically weighted regression analysis (GWRA) compared to a standard regression (Fotheringham *et al.* 2002).

In GWPCA (LWPCA) a PCA analysis is carried out in a geographic-space (attribute-space) neighborhood for each observation, restricting the calculations to that neighborhood where homogeneity of the covariance is assumed. The size of the vicinity over which a local PCA might apply is controlled by the bandwidth. Small bandwidth values lead to more rapid variation in the results, whereas very large bandwidths give subspaces increasingly close to the standard PCA solution (Demsar *et al.* 2013). As a result, a number of principal components analyses equal to the number of observations is conducted. A detailed analysis of the results can provide information concerning the internal structure of the data (Lloyd 2012, Kumar *et al.* 2012).

Although results of GWPCA such as the loadings or the percentage of variance explained for each component can show variability in the data structure, a complete analysis should be accompanied by a statistical contrast in order to establish if the variability of the covariance is significant from a statistical point of view. Otherwise, the GWPCA would not be justified. Harris *et al.* (2011) propose a randomization Monte Carlo test for evaluating the significance of eigenvalue variability, following a similar procedure to that used in GWR to test if local regression parameters vary significantly across space (Brunsdon *et al.* 1998). The idea is to determine the standard deviation of each local eigenvalue in a rank distribution of the standard deviations obtained applying GWPCA to each randomized data set. This kind of test has been implemented in the R (Ihaka and Gentleman 1996) GWmodel package (Gollini *et al.* 2015). In this paper we propose a different approach to tackle the problem of testing spatial heterogeneity in the data based on defining a statistic from the new variables obtained after applying a standard PCA to the raw data. The proposed statistic uses the eigenvectors of the covariance matrix and estimates its level of significance from the distribution function of that statistic obtained by means of Monte Carlo simulation.

This paper is organized as follows: Section 2 introduces the geographically weighted principal components analysis including the problem of bandwidth selection. In Section 3 a new statistical test to check spatial heterogeneity on the data is proposed. In Section 4 a simulation study was conducted to assess the validity of the proposed statistical test. In Section 5 the exposed methodology is applied to a real problem whose objective is to find spatial distribution patterns in a set of soil pollutants. Finally, the conclusions of the paper are reported.

2. Geographically Weighted Principal Components Analysis

Let us consider that $\mathbf{x}$ is a vector of p random variables with a matrix of covariances Σ. By definition Σ is a positive semi-definite matrix, and therefore it is possible to perform an eigen decomposition according to

$$\Sigma = \mathbf{P}\Lambda\mathbf{P}^t \tag{1}$$

where Λ is the diagonal matrix of eigenvalues of Σ and $\mathbf{P}$ is an orthogonal projection called loading matrix, whose k_{th} column is the k_{th} eigenvector of Σ. We have directly that $\mathbf{P}^t\Sigma\mathbf{P} = \mathbf{\Lambda}$ (for a comprehensive text regarding principal component analysis see Jolliffe (2002). Usually the elements of $\mathbf{\Lambda}$, the eigenvalues, are in descending order, so that

$$\Lambda = diag(\lambda_1, \ldots, \lambda_p) \quad \text{with} \quad \lambda_1 \geq \lambda_2 \ldots \geq \lambda_p \geq 0 \tag{2}$$

In this way, the columns of $\mathbf{P}$ are the directions of maximum variance in the data, with the first column representing the direction of maximum variance, the second column the direction of the next largest variance, and so on. These directions correspond to the eigenvectors of either the data covariance or correlation matrix Σ. Principal component analysis maps the original $\mathbf{x}$ onto a new orthogonal space following the transformation

$$\mathbf{z} = \mathbf{P^t x} \tag{3}$$

so that the new axes are oriented in directions of largest variance in the data.

We are concerned with a spatial study where the standard PCA is replaced with a spatial variant approach. In these situations, the vector of variables $\mathbf{x}$ has a spatial location, given by $\mathbf{s} = (s_1, s_2)$ (two-dimensional locations, for example) (Demsar *et al.* 2013, Harris *et al.* 2011). In this case, the spatial data may not be well described by a global model of PCs but there are localized regions in the attribute data space where a suitably localized set of PCs provide a better description (local models). That is, in different parts of the data space, a different set of PCs is needed. This heterogeneity can be modelled by a covariance matrix depending on the spatial positions. More explicitly, the covariance matrix at $\mathbf{s}$ can be expressed as

$$\Sigma(\mathbf{s}) = \mathbf{P}(\mathbf{s})\Lambda(\mathbf{s})\mathbf{P}(\mathbf{s})^t \tag{4}$$

where $\Lambda(\mathbf{s})$ is the diagonal matrix of eigenvalues of $\Sigma(\mathbf{s})$, and $\mathbf{P}(\mathbf{s})$ is an orthogonal projection matrix verifying $\mathbf{P}(\mathbf{s})^t\Sigma(\mathbf{s})\mathbf{P}(\mathbf{s}) = \Lambda(\mathbf{s})$.

In the next section an algorithm that allows to obtain locally estimates of $\Sigma(\mathbf{s})$ is exposed. This technique uses a moving window weighting approach in the data space where PCs are found in the vicinity of some target location $\mathbf{s}$ in the data space. All neighboring observations are weighted according to some distance-decay kernel function that quantifies the spatial dependence between the observed variables. The size of the window is controlled by the bandwidth, a parameter of that kernel function. Small bandwidths lead to more rapid spatial variation in the results while large bandwidths yield results increasingly close to the universal model solution (Gollini *et al.* 2015).

2.1. *Nonparametric estimation algorithm*

Given a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of $\mathbf{x}$ with associate spatial positions $\mathbf{s}_1, \ldots, \mathbf{s}_n$, the estimated variance matrix $\hat{\Sigma}$ in the spatial position (s_1, s_2) is obtained as follows:

$$\hat{\Sigma}(s_1, s_2) = \mathbf{X^t W}(s_1, s_2)\mathbf{X} \tag{5}$$

where $\mathbf{X}$ is the data matrix with n rows representing the observation and p columns representing the variables. We assumed that the columns of $\mathbf{X}$ have been standardised with zero mean and unit variance, i.e, PCA is based on correlation matrix. Moreover, $\mathbf{W}(s_1, s_2) = \text{diag}\{W_1(s_1, s_2), \ldots, W_n(s_1, s_2)\}$ is a diagonal matrix of geographic weights. Finally, the estimated matrices $\hat{\mathbf{P}}(s_1, s_2)$ and $\hat{\Lambda}(\mathbf{s})$ originating the eigen decomposition

$$\hat{\Sigma}(\mathbf{s}) = \hat{\mathbf{P}}(s_1, s_2)\hat{\Lambda}(s_1, s_2)\hat{\mathbf{P}}^t(s_1, s_2) \tag{6}$$

are obtained as in the standard PCA.

In this paper, we have used the kernel weights given by

$$W_i(s_1, s_2) = \frac{w_i(s_1, s_2)}{\sum_{j=1}^{n} w_j(s_1, s_2)}; w_i(s_1, s_2) = \exp\left(-\frac{\sqrt{(s_{i1} - s_1)^2 + (s_{i2} - s_2)^2}}{h}\right)$$

Note that $W_i(s_1, s_2)$ is a weighted function depending on the euclidean distance between (s_1, s_2) and (s_{i1}, s_{i2}) and, in addition, contains a smoothing positive parameter h usually called smoothing parameter or bandwidth. Another type of kernel apart from the Gaussian, such as exponential or bisquare, may also have been tested. An interesting discussion on density and regression kernel estimation can be found in Wand and Jones (1995).

Because of the definition of W, the observation close to $\mathbf{s} = (s_1, s_2)$ has more influence on the estimate $\hat{\Sigma}(\mathbf{s})$ than those farther away. The amount of relative influence is controlled by the bandwidth h. On the one hand, if h is small the resulting estimate $\hat{\Sigma}(\mathbf{s})$ heavily depends on those observations that are closest to (s_1, s_2) and tends to yield a more changeful estimate. On the other hand, if the bandwidth is too large $W_i(s_1, s_2) \to n^{-1}, i = 1, \ldots, n$ the estimates $\hat{\Sigma}(\mathbf{s})$ will not depend on spatial location (s_1, s_2) and, consequently, will not adjust to the real shape of the true $\Sigma(\mathbf{s})$. This shows the importance of arranging a tool to help to choose the most appropriate smoothing bandwidth. Different schemes can be considered in bandwidth selection: a fixed scheme, using a constant bandwidth at each spatial location or an adaptive scheme, that allows different bandwidth values at each location. Adaptive schemes involve estimating the bandwidth from the k nearest neighbors, so the bandwidth can be different at each location depending on the distance to those neighbors. The number of nearest neighbors can be set directly or estimated using cross-validation, as we have done in this work following Harris *et al.* (2011) and Gollini *et al.* (2015).

2.2. *Bandwidth selection*

The specification of the bandwidth h is very important as shown above, and is a problem that is yet to be solved, not only in GWPCA, but also in other mathematical methods concerning local estimations, such as geographically weighted regression (GWR) (Fotheringham *et al.* 2002). Following Harris *et al.* (2011), we used the cross-validation automatic selection bandwidth based on the approximations obtained with the first q principal components. The information in $\mathbf{X}$ can be approximated by a small number of PCs, q, where

$q < p$, while still explaining most of the variance in the data; that is, when Λ only has a small number of large eigenvalues and many small ones.

Denoting $\mathbf{P}_q$ as the matrix containing the first q columns of $\mathbf{P}$, then the corresponding scores are given by $\mathbf{Z}_q = \mathbf{XP_q}$ and the proportion of the total variance explained is $\sum_{i=1}^{q} \lambda_i / trace(\Lambda)$. Therefore an approximation of $\mathbf{X}$ is given by $\hat{\mathbf{X}} = \mathbf{X}\mathbf{P}_q\mathbf{P}_q^t$.

In a context of local PCA, the bandwidth h can be selected by means of cross-validation, that is, minimizing

$$CV(h) = \sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i^{(-i)}|| \tag{7}$$

where $||\cdot||$ represents the euclidean distance and

$$\hat{\mathbf{x}}_i^{(-i)} = \hat{\mathbf{P}}_{q,(-i)}(\mathbf{s}_i)\hat{\mathbf{P}}_{q,(-i)}^t(\mathbf{s}_i)\mathbf{x}_i$$

being $\hat{\mathbf{P}}_{q,(-i)}(\mathbf{s}_i)$ the matrix containing the first q columns of $\hat{\mathbf{P}}(\mathbf{s}_i)$ leaving out the i-sample data $(\mathbf{s}_i, \mathbf{x}_i)$.

Note that to use this method we have to choose a priori the number q of principal components to retain. Different bandwidths for each number $q = 1, \ldots, p-1$ of the principal components to be retained may be obtained. If $q = p$, i.e when all components are retained, cross-validation is not a valid procedure because $||\mathbf{x}_i - \hat{\mathbf{x}}_i^{(-i)}|| = 0$. Therefore, the procedure can be used as an exploratory selector, or a reasonable starting point to help us select the bandwidth. Further research is needed to establish the adequate selection of the bandwidth parameter, but our recommendation is to conduct the study over a grid of values.

3. Testing spatial structure

In order to justify the use of GWPCA instead of standard PCA, from a statistical point of view, a hypothesis test should be performed previously. We are interested in the problem of testing if the hypothesis of the variability of the covariance is significant over the spatial position. In particular, in this section we shall consider the development of a hypothesis test for

$$H_0 : \Sigma(\mathbf{s}) = \Sigma \quad \forall \mathbf{s} \quad \text{against} \quad H_1 : \Sigma(\mathbf{s}) \text{ not all the same } \forall \mathbf{s} \tag{8}$$

An Omnibus test to detect any departure from the null hypothesis would be desirable. A viable alternative, broad-spectrum test is given in Harris *et al.* (2011). They propose measuring the variability of $\Sigma(\mathbf{s})$ using the standard deviation of a local eigenvalue. The statistic proposed is given by

$$T_1 = n^{-1} \sum_{i=1}^{n} \left(\hat{\lambda}_1(\mathbf{s}_i) - \bar{\lambda}_1 \right)^2 \tag{9}$$

where $\hat{\lambda}_1(\mathbf{s}_i)$ is the first eigenvalue of the matrix $\hat{\Lambda}(\mathbf{s}_i)$, and $\bar{\lambda} = n^{-1} \sum_{i=1}^{n} \hat{\lambda}_1(\mathbf{s}_i)$. Clearly, under non-geographical spatial variability on the covariance structure, the value of T_1 should be close to zero.

In this paper, we propose an alternative statistic based on eigenvectors instead of eigenvalues. We consider the transformed data $\mathbf{z}_1, \ldots, \mathbf{z}_n$ with $\mathbf{z}_i = \mathbf{P^t x}_i$ being $\mathbf{P}$ the projection matrix of $\mathbf{X}$ in a standard PCA. Denoting the covariance matrix of $\mathbf{z}$ as $\Sigma_{\mathbf{z}}(\mathbf{s})$ the following spectral decomposition is obtained

$$\Sigma_{\mathbf{z}}(\mathbf{s}) = \mathbf{P}_{\mathbf{z}}(\mathbf{s})\Lambda_{\mathbf{z}}(\mathbf{s})\mathbf{P}^t_{\mathbf{z}}(\mathbf{s})$$

with $\mathbf{P}_{\mathbf{z}}(\mathbf{s})$ an orthogonal projection and $\Lambda_{\mathbf{z}}(\mathbf{s})$ diagonal matrix of corresponding eigenvalues.

Clearly, under H_0 the matrix of covariances of $\Sigma_z(\mathbf{s})$ has no spatial structure, and therefore the matrix $\mathbf{P}_{\mathbf{z}}$ is given by

$$\mathbf{P}_{\mathbf{z}}(\mathbf{s}) = \mathbf{P}_{\mathbf{z}} = \mathrm{diag}(1, \ldots, 1) \tag{10}$$

The statistic proposed below, that we denote as T_2, is based on the difference between estimated matrix $\hat{\mathbf{P}}_{\mathbf{z}}(\mathbf{s})$ and the expected matrix under H_0 given in (10).

The process to compute T_2 is the following:

- Firstly, using the sample data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ obtain the estimated covariance matrix $\hat{\Sigma} = n^{-1}\mathbf{X}^t\mathbf{X}$ and the corresponding spectral decomposition $\hat{\Sigma} = \hat{\mathbf{P}}\hat{\Lambda}\hat{\mathbf{P}}^t$.
- Secondly, using the projected data $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_n$ and associated spatial positions $\mathbf{s}_1, \ldots, \mathbf{s}_n$, obtain the estimated decomposition

$$\hat{\Sigma}_{\mathbf{z}}(\mathbf{s}) = \hat{\mathbf{P}}_{\mathbf{z}}(\mathbf{s})\hat{\Lambda}_{\mathbf{z}}(\mathbf{s})\hat{\mathbf{P}}^t_{\mathbf{z}}(\mathbf{s})$$

- Finally, compute the proposed statistic

$$T_2 = n^{-1}\sum_{i=1}^{n}\left(\hat{\mathbf{P}}^{(1,1)}_{\mathbf{z}}(\mathbf{s}_i) - 1\right)^2 \tag{11}$$

where $\hat{\mathbf{P}}^{(1,1)}_{\mathbf{z}}(\mathbf{s}_i)$ represent the $(1,1)$th element of the matrix $\hat{\mathbf{P}}_{\mathbf{z}}(\mathbf{s}_i)$. Under non-spatial structure, the first column (PC_1) of $\hat{\mathbf{P}}_{\mathbf{z}}(\mathbf{s})$ will be close to vector $(1, 0, \ldots, 0)$, the second column (PC_2) close to $(0, 1, \ldots, 0)$, and so on.

Note that under the null hypothesis the value of T - both for T_1 and T_2- should be close to zero. For a given significance α, the decision rule based on T consists of rejecting the null hypothesis if $T > T^{\alpha}$, where T^{α} represents the $(1-\alpha)$-percentile of T.

Unfortunately, the theory for determining such percentiles is not closed. Brunsdon *et al.* (1998) proposed using Monte Carlo techniques (Hope 1968).

Explicitly, the procedure is as follows:

For $b = 1, \ldots, B$ (e.g. $B = 1000$),

Step 1: A re-sample data $\mathbf{s}^{*,b}_1, \ldots, \mathbf{s}^{*,b}_n$ is obtained as a random permutation of the original data $\mathbf{s}_1, \ldots, \mathbf{s}_n$.

Step 2: Using the original data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and the sampled positions $\mathbf{s}^{*,b}_1, \ldots, \mathbf{s}^{*,b}_n$ obtained in **Step 1**, compute the test statistic T^{*b} as in (9).

Step 3: Finally, the null distribution of T is approximated by the empirical distribution of the values $T^{*1}, \ldots, T^{*B}$. Therefore, the test rule consists of rejecting the null

hypothesis if $T > \hat{T}^{\alpha}$, where $\hat{T}^{\alpha}$ is the empirical $(1-\alpha)$-percentile of values $T^{*1}, \ldots, T^{*B}$.

As will be shown below, the power of this hypothesis test obtained with this procedure using statistic T_1, in our simulation study, is quite poor, at least in comparison with the power obtained with the proposed statistic T_2.

4. Simulation Study

This section reports the results of a simulation study to assess the validity of the testing procedures exposed above. Both T_1 and T_2 statistics are compared.

To perform the simulation, a thousand independent samples $\{\mathbf{s}_i, \mathbf{x}_i\}_{i=1}^{n}$ were generated where each $\mathbf{s_i}$ was drawn from an independent bivariate uniform $U[-2,2] \times U[-2,2]$, and $\mathbf{x}_i$ is a $p-$dimensional vector drawn from a zero mean Gaussian distribution with covariance matrix given by

$$\Sigma(\mathbf{s}_i) = \begin{pmatrix} 1 & \rho(\mathbf{s}_i) & \ldots & \rho(\mathbf{s}_i) \\ \rho(\mathbf{s}_i) & 1 & \ldots & \rho(\mathbf{s}_i) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(\mathbf{s}_i) & \rho(\mathbf{s}_i) & \ldots & 1 \end{pmatrix} \tag{12}$$

Two simulation stages were considered: in stage (i) we assumed that $\rho(\mathbf{S}_i) = \rho$, a constant positive value less than or equal to 1. In fact, three different values of ρ were considered. In stage (ii) it was considered that

$$\rho(\mathbf{s}_i) = \min\left(a\left|s_{i1}^2 - s_{i2}^2\right|, 0.99\right) \tag{13}$$

being $a > 0$ a constant. In that way the correlation matrix depends on each sample location and the spatial heterogeneity can be easily simulated. Also, the statistical power of each test can be analyzed in terms of the parameter a.

Note that under stage (i) H_0 should never be rejected. However, under stage (ii) the value $a = 0$ corresponds to the null hypothesis H_0, but as the value of a rises the model departs from that hypothesis. High a values correspond to high correlations among the variables (higher spatial dependence). Therefore, the dependence of the variables comes from the spatial locations of the data. However, the marginal distributions of the variables do not have spatial dependence, although this does not invalidate our experiment since it is focused on testing if there is spatial heterogeneity, that is, if correlation between variables changes throughout the study area. In addition, defining the elements of the correlation matrix using a mathematical expression instead of determining it from the data, allows us to study the statistical power of each test by simply modifying the parameter a according to (13).

To determine the critical values of the test statistics, we applied bootstrap as described above. Both the estimated type I and type II errors were calculated on the basis of 1000 simulation runs. It must be remembered that type I error provides the decision of rejecting the null hypothesis when it is true, and that it is fixed a priori by the significance level (probability of assuming a type I error). On the contrary, the statistical power (1

- probability of making a type II error) of the test is the capacity of rejecting the null hypothesis when it is false.

Estimated type I error (in %) under stage (i) for testing H_0 for different sample sizes ($n = 100$, $n = 200$ and $n = 500$), bandwidths ($h = 5$, $h = 10$ and $h = 15$), correlations ($\rho = 0.5$, $\rho = 0.8$ and $\rho = 0.95$) and number of covariates ($p = 5$ and $p = 10$) are shown in Table 1. As can be appreciated, the test performed reasonably well, with almost all the values holding the level and several coming quite close to the nominal level. Moreover, the results are similar for the different combinations of (n, ρ, h) considered.

Table 1.: Estimated type I error for testing H_0 in stage (i) for $p = 5$ and $p = 10$ and different values of the sample size n and the bandwidth h.

				$\rho = 0.5$			$\rho = 0.8$			$\rho = 0.95$		
			test	1%	5%	10%	1%	5%	10%	1%	5%	10%
p=5	n=100	$h = 0.05$	T_1	0.9	4.7	9.7	0.5	4.8	9.2	0.8	4.0	8.6
			T_2	1.2	4.9	9.2	0.7	4.7	10.2	0.6	4.8	9.2
		$h = 0.10$	T_1	0.9	4.7	9.7	0.5	4.8	9.2	0.8	4.0	8.6
			T_2	1.2	4.9	9.2	0.7	4.7	10.2	0.6	4.8	9.2
		$h = 0.15$	T_1	1.4	4.5	10.1	1.0	5.1	9.6	1.1	5.2	10.6
			T_2	0.6	4.5	10.1	0.7	3.6	8.2	1.2	5.2	10.5
	n=500	$h = 0.05$	T_1	1.1	4.5	10.8	1.2	5.7	10.9	1.3	4.9	10.4
			T_2	1.2	5.9	10.6	1.2	5.2	11.3	1.0	5.6	11.1
		$h = 0.10$	T_1	0.8	4.5	9.4	0.9	4.7	9.0	0.9	5	10.7
			T_2	1.2	5.2	8.9	1.0	5.0	10.5	0.7	5.0	9.9
		$h = 0.15$	T_1	0.9	5.3	10.7	0.9	4.3	9.7	1.0	6.1	10.8
			T_2	1.2	5.2	10.2	0.9	4.8	8.6	0.6	4.8	10
	n=1000	$h = 0.05$	T_1	1.6	5.2	10.2	1.3	5.7	10.7	1.7	5.5	11.5
			T_2	0.9	4.6	9.0	1.4	4.9	10.0	1.3	4.9	10.8
		$h = 0.10$	T_1	1.0	5.5	10.9	0.9	5.2	10.3	1.2	4.6	7.8
			T_2	0.8	5.4	10.8	0.9	4.3	9.7	1.1	4.9	11.1
		$h = 0.15$	T_1	0.9	4.5	10.0	1.0	5.8	10.4	0.8	4.2	9.8
			T_2	1.2	5.3	10.6	0.8	5.0	10.0	0.8	4.6	9.3
p=10	n=100	$h = 0.05$	T_1	0.7	4.7	10.8	1.1	6.8	11.7	0.7	4.1	7.8
			T_2	1.1	5.1	10.2	0.8	5.3	10.4	0.3	4.0	9.7
		$h = 0.10$	T_1	0.7	4.7	10.8	1.1	6.8	11.7	0.7	4.1	7.8
			T_2	1.1	5.1	10.2	0.8	5.3	10.4	0.3	4.0	9.7
		$h = 0.15$	T_1	0.8	5.7	11.2	1.2	4.9	9.6	0.8	5	9.9
			T_2	0.8	4.6	11.2	1.3	4.9	9.8	0.8	4.7	11
	n=500	$h = 0.05$	T_1	0.5	4.9	10.5	0.6	5.5	10.1	0.8	4.4	10
			T_2	0.9	4.8	10.8	1.5	5.8	11.3	1.2	4.5	10.5
		$h = 0.10$	T_1	0.9	5.2	11.4	0.6	4.6	9.0	0.6	6.3	11.1
			T_2	0.6	5.1	10.8	1.4	5.7	10.6	1.0	5.6	12.2
		$h = 0.15$	T_1	1.0	4.9	10.2	0.5	4.0	8.9	0.3	5.5	10.1
			T_2	0.6	3.8	9.8	1.1	5.2	9.3	0.9	4.6	10.0
	n=1000	$h = 0.05$	T_1	0.8	4.1	8.3	0.8	5	10.3	1.9	5.9	9.6
			T_2	0.7	4.1	8.9	0.5	2.8	9.0	0.9	5.3	10.3
		$h = 0.10$	T_1	0.6	5.3	9.8	0.5	4.1	8.7	0.9	5.2	9.7
			T_2	0.5	4.7	11	1.0	5.1	11.2	1.2	4.7	10.9
		$h = 0.15$	T_1	1.2	4.4	9.3	1.1	6.0	11.0	1.5	5.3	11.7
			T_2	1.2	5.1	10.1	0.9	5.1	10.2	1.2	5.0	10.4

We also studied the performance of the alternatives, as a function of a in stage (ii). Power values are shown in Figure 1 and Figure 2. It is easy to appreciate in Figure 1 that statistic T_2 produces satisfactory power curves, with the probability of rejection rising in response to any increase in the value of the constant a. Furthermore, as was expected, the power increases with the sample size. In general the proposed test T_2 presents better

power curves than those corresponding to test T_1.

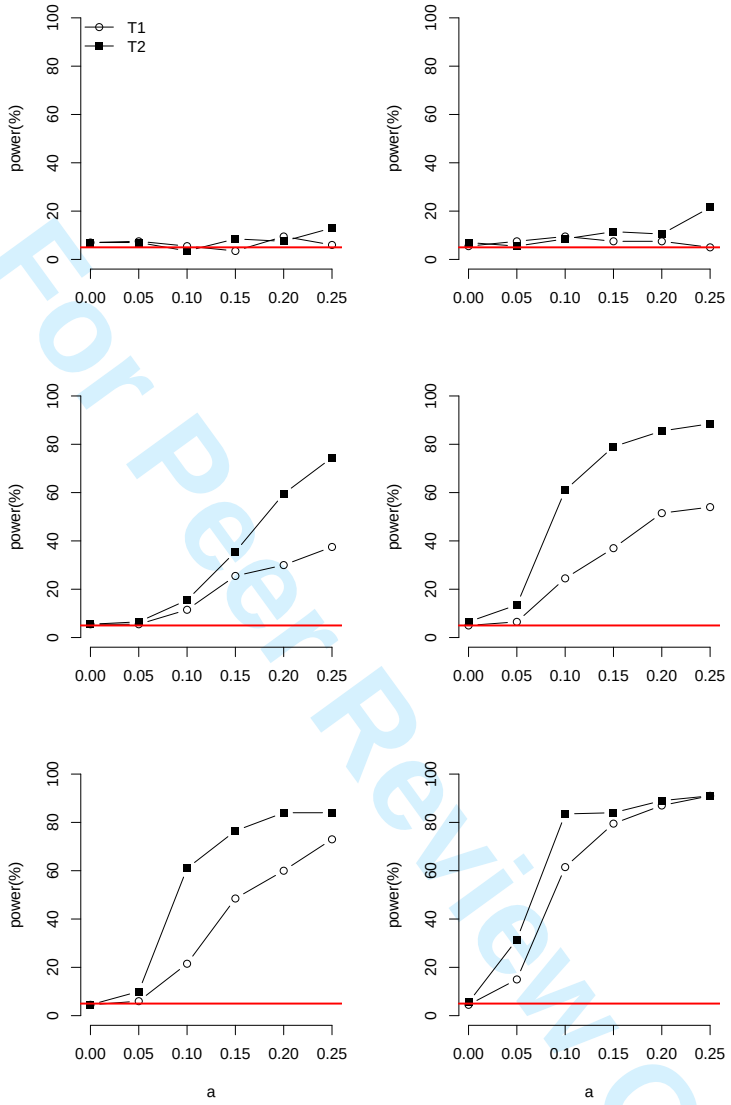

Figure 1.: Percentage of rejections for test statistics T_1 and T_2 on increasing a for nominal level of 5%, number of covariates $p = 5$ and $p = 10$(left and right panel, respectively). Upper panel: rejections for sample size $n = 100$. Middle panel: rejections for sample size $n = 500$. Lower panel: rejections for sample size $n = 1000$.

The power curves in Figure 1 were obtained for $h = 10$; however, results strongly depend on this paramenter. Figure 2 represents the power curves for different values of a corresponding to test T_2 and different values of h. Moreover, Figure 3 depicts the power of the test T_2 for two different values of a ($a = 0.1$ and $a = 0.2$). As can be appreciated in those figures, the power curve has the typical inverted U shape of smoothing techniques. Therefore a bandwidth which is too small or too large produces poor power. For intermediate bandwidth values the power increases with the value of h.

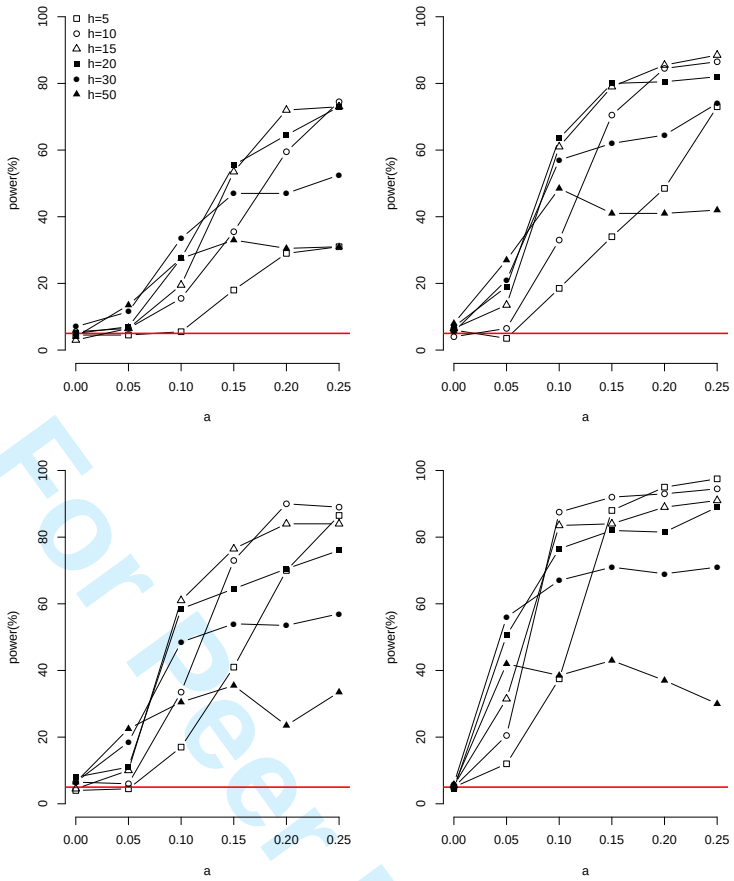

Figure 2.: Percentage of rejections for the test based on statistic T_2 for increasing values of a, nominal level 5% and number of covariates $p = 5$ and $p = 10$(left and right panel, respectively). Upper panel: rejections for sample size $n = 100$. Lower panel: rejections for sample size $n = 1000$.

5. Application to real data: soil contamination analysis

5.1. *GWPCA*

The proposed methodology to evaluate spatial heterogeneity in spatial data was applied to the analysis of the spatial distribution of a set of samples of soil pollutants in a specific area. This area is located in Avilés (Northern Spain) and corresponds to a very industrialized sea port with dense maritime traffic and several chemical and metallurgical (steel, Zn and Al) production plants. High levels of pollution and the presence of nearby beaches and populated urban areas have led this place to be the subject of several studies aimed at analyzing the sources of pollution, the distribution of the pollutants and possible solutions to the pollution problem (Berciano *et al.* 1989, Gallego *et al.* 2002).

Four sub-areas can be distinguished in the study according to their geochemical and sedimentary characteristics (Figure 4); namely: Salinas-El Espartal, a protected eolian dune system with slightly developed soils that are highly deteriorated due to industrial activities; Llodero Cove, a zone with abundant intertidal mud flats and rich in organic matter where aluminium and steel factories are located (Flor-Blanco *et al.* 2013); the

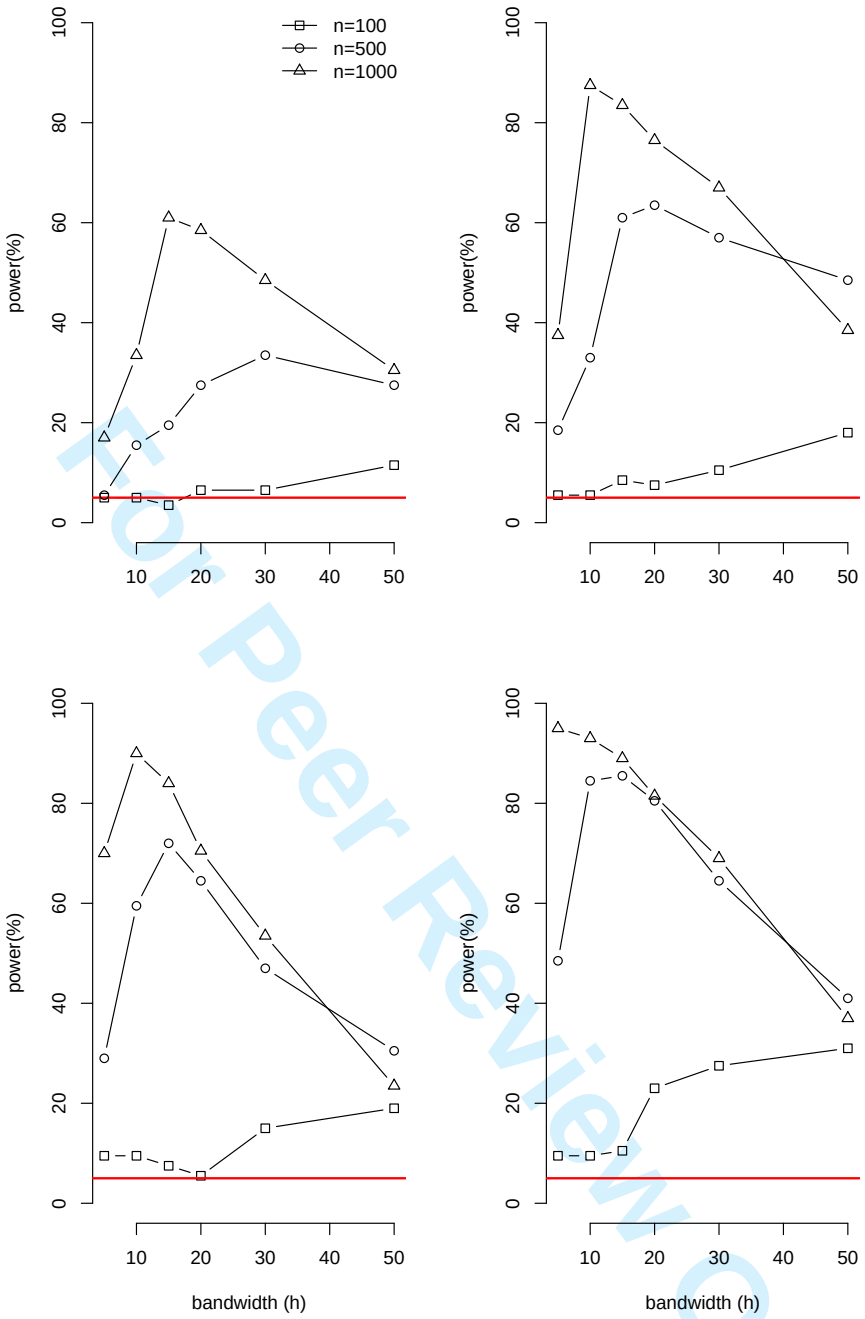

Figure 3.: Percentage of rejections for statistic test T_2 on increasing h for nominal level of 5% and number of covariates $p = 5$ and $p = 10$ (left and right panel, respectively). Upper panel: rejections for $a = 0.1$. Lower panel: rejections for $a = 0.2$.

fluviomarine terraces of the estuary margins; and finally, Xagó Beach, a zone with less anthropogenic disturbances and analogous lithology, thus taken as natural background.

Samples were collected in duplicate with a modified Van Veen grab sampler. The distance between successive points was approximately 10 m, and between duplicated points approximately 2 m.

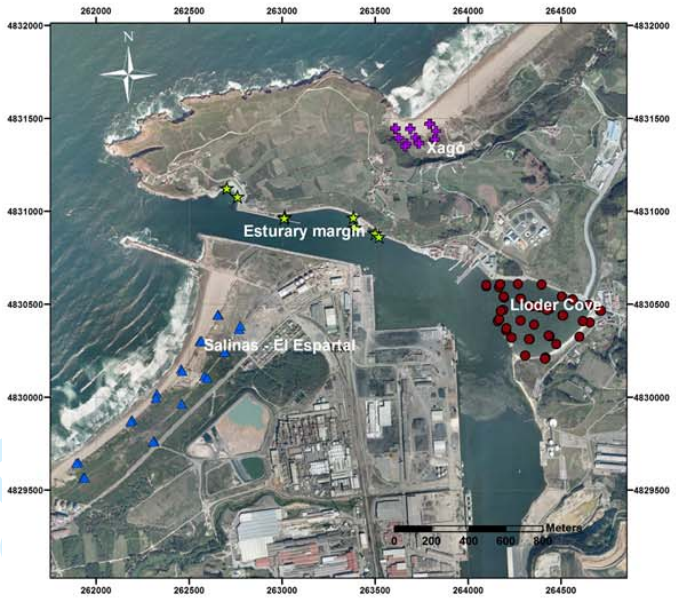

Figure 4.: Location of the samples overlaid on an orthophotography of the study area (Coordinate system: ETRS1989 UTM Zone 30N).

Particle size characterization was performed by means of laser diffraction spectroscopy (LS 13-320 MW model -Beckman Inc. Coulter), after dispersion with sodium hexametaphosphate and sodium carbonate and elimination of the organic matter with hydrogen peroxide (Gee and Bauder, 1996). In our analysis the sample size is 212 and 18 pollution variables: Cu, Pb, Zn, Ag, Ni, Co, Mn, Fe, As, Sr, Cd, V, Ca, La, Cr, Mg, Al, Na, K, S.

Table 2 represents the results of applying a global PCA to the data. Note that the cumulative proportion of variance explained for the first five global components is 0.85. The loadings for each component provide information regarding the importance of each pollutant in that component. Ni, Co, Fe and V are the elements with the highest loadings in the first principal component, so this component can be partially associated to industrial and urban acitivities. For instance, V-Ni associations are usually related to the combustion of fossil fuel. The elements with the highest absolute loadings in the second principal component are Sr, Ca and Mg, which are related to the alteration of carbonates. Accordingly, the analysis of the loadings for each component suggests that there must be some relationship between the different pollutants that cannot be observed physically. However, this analysis does not give any information about possible changes in this relationship throughout the study area.

Table 2.: Results of a global PCA.

	PC1	PC2	PC3	PC4	PC5
Standard deviation	2.8278	1.7080	1.6164	1.4249	1.2419
Proportion of variance	0.3998	0.1459	0.1306	0.1015	0.0771
Cumulative proportion	0.3998	0.5457	0.6763	0.7778	0.8549
Loadings					
Cu	0.21	0.04	**-0.45**	0.17	-0.06
Pb	0.22	0.10	**-0.41**	0.18	-0.12
Zn	0.20	0.04	-0.30	**-0.35**	0.16
Ag	0.11	-0.06	-0.31	**-0.40**	0.25
Ni	**0.31**	0.12	-0.14	0.23	-0.14
Co	**0.33**	0.13	0.05	0.07	-0.13
Mn	0.27	0.10	0.32	-0.10	-0.11
Fe	**0.31**	0.13	0.14	0.10	-0.14
As	0.15	-0.03	-0.11	-0.33	-0.00
Sr	0.02	**-0.43**	-0.11	-0.05	-0.37
Cd	0.19	0.03	-0.06	**-0.37**	0.25
V	**0.33**	-0.00	0.17	0.06	0.01
Ca	0.01	**-0.49**	-0.10	-0.08	-0.37
La	0.27	0.3	0.33	-0.11	-0.06
Cr	0.30	0.04	-0.15	0.30	0.01
Mg	0.06	**-0.45**	-0.01	-0.00	-0.18
Al	0.27	-0.19	0.22	-0.14	0.08
Na	0.02	-0.28	0.03	**0.37**	**0.47**
K	0.25	-0.25	0.22	-0.04	0.16
S	0.12	-0.33	-0.04	0.24	**0.45**

In order to test the existence of spatial heterogeneity, the statistical tests of Section 3 were applied to the dataset. Firstly, we compute the cross-validation function as a function of h. As can be seen in Figure 5 the minimum value of the CV error is obtained for the bandwidth $h = 0.18$.

To evaluate the influence of the bandwidth on the results, the GWPCA was conducted for several values of h. Table 3 shows the p-value obtained using T_1 and T_2 statistics for different bandwidths. It can be observed that statistical significance (p-value< 0.05) was obtained for all the bandwidths below 0.30 (30% of the dataset) in both cases. When $h > 0.30$ (again h is measured as a percentage of the total number of data) the statistical significance is lost as higher p-values are obtained. This is consistent with the simulation study previously conducted, that showed that high values of h produce hypothesis contrasts with very low power. A bandwidth value greater than 0.18 gives a larger cross-validation score, therefore $h = 0.18$ was selected to carry out the GWPCA. Anyway, it would be advisable to check the results of the statistical test for spatial heterogeneity with different bandwidth values close to those obtained with cross-validation.

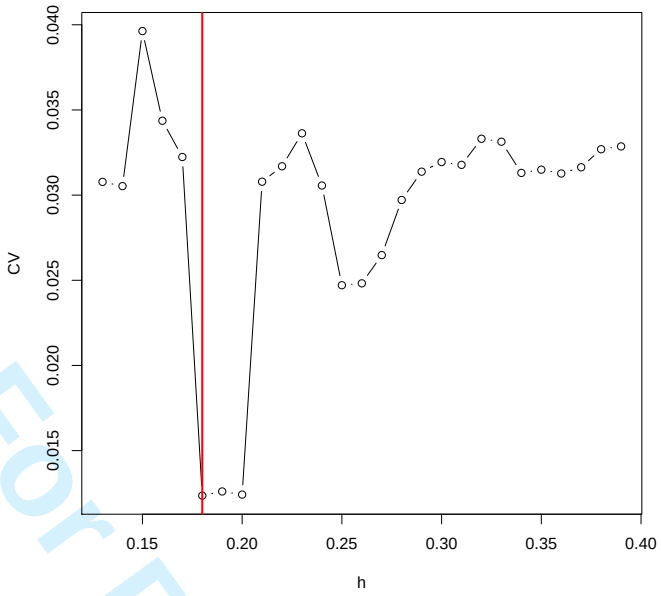

Figure 5.: Cross validation function as a function of h (with four retained components)

Table 3.: p-value obtained for different bandwidth sizes h.

h	T_1	T_2
0.01	0.00	0.00
0.05	0.00	0.00
0.10	0.00	0.00
0.15	0.00	0.00
0.20	**0.00**	**0.00**
0.25	0.01	0.00
0.30	0.09	0.06
0.35	0.58	0.54
0.40	0.92	0.76
0.45	0.94	0.96
0.50	0.88	0.92
0.55	0.83	0.83
0.60	0.74	0.83

Figure 6 shows the spatial distribution of the percentage of variance explained for the first four principal components. The fact that the variance changes throughout the study area also suggests the advantage of using GWPCA against global PCA. Note that the percentage of variance explained for the first four local components exceeds the percentage of variance explained for the first five components in the global PCA, in some points.

One of the main differences observed between the introduced mathematical procedure and the traditional PCA is that it offers the possibility of making not only a conjunct interpretation for all data, but also as many analyses as there are data, according to their location. Figure 7 shows the spatial distribution of the winning variables for the first four geographically weighted components analyzed. Winning variables are those with

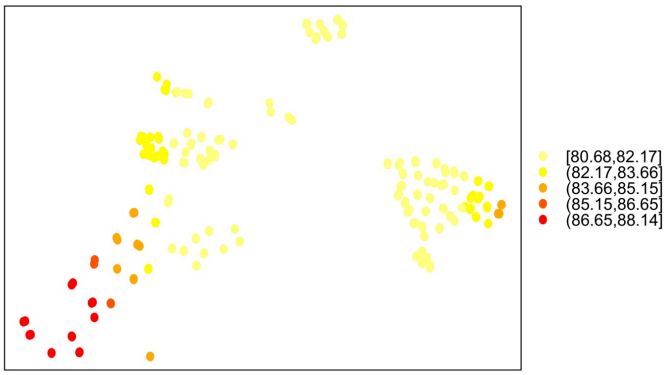

Figure 6.: Percentage of local variance for the first four local components.

the highest absolute local loading in the corresponding component. Then they are related to the importance of those variables in each component and provide useful information regarding the pattern of spatial distribution of the pollutants. Note that it is possible to appreciate the clustering of some of the pollutants in the sub-areas considered.

Figure 7.: Winning variables for the first four geographically weighted components (GWPCs).

Accordingly, the explanation that can be offered for GWPC1 is the same for all the sub-areas, and evidences a strong correlation within the elements that represent the contamination (chalcophiles such as Zn, Pb, Cd and Cu). This geochemical association has been widely described in the scientific literature, e.g. Burton *et al.* (2005), and is typically connected with the ores used in Zn production (sphalerite and galena). Therefore,

we suggest that the origin of this association lies in the dust generated by the mineral transport and storage taking place in the port as well as in the particulate emissions from Zn metallurgy e.g. (Li *et al.* 2006, Mattielli *et al.* 2009). Copper emissions also seem likely to have occurred as a consequence of the same source. Among the major sources for Cd contamination we also suggest atmospheric deposition but in this case enhanced by the smelting plants in the surrounding areas as well as the burning of fossil fuels.

Similarly, the analysis of the spatial distribution of the winning variable in GWPC2 allows us to conclude that this component is connected with lithogenic elements such as Al, K and La, reflecting the siliciclastic materials from the geology of the area Flor-Blanco *et al.* (2013), as well as elements of biogenic origin or alteration of the carbonates (for instance Ca and Mg from shells). As in the previous component, this appreciation is similar irrespective of the considered sub-areas.

Likewise, GWPC3 represents, in general terms, elements that are not present in the lithology of the zone. A good example of this statament is Na, an element of non-lithogenic origin which plays a major role in the component, but with influence focused on the zone of more marine influence, namely the Llodero Cove area.

On the other hand, GWPC4 represents a group of mixed nature in which anthropogenic but mainly natural elements are present. In this respect, the natural component appears to dominate the factor; however, in the dune area of Salinas an important association between As/Cd was observed and could be attributed to metallo-organic complexing of both elements in the abundant organic matter present in the topsoil.

6. Conclusions

GWPCA requires a previous study in order to prove the existence of spatial heterogeneity in the data. The main contribution of our work is to propose a statistical contrast to solve this problem which, according to the simulation study performed, leads to a better statistical power than an extended method based on calculating the eigenvalues of the covariance matrix. Therefore, there is less chance of making a type II error, that is, of considering spatial homogeneity in the data when there is none in reality. Otherwise, we could decide not to apply GWPCA when it is advisable. The proposed method is useful for those users interested in testing for spatial heterogeneity, such as geographers, biologists or geologists, among others.

First, we calculate PCA scores from raw data and then the eigenvectors of the covariance matrix, instead of the eigenvalues as proposed by other authors, are determined. Despite the good performance of our method, we have also proved that the results obtained using both methods strongly depend on the bandwidth values, that is, on the size of the neighborhood used to perform the principal components analysis locally. As the determination of the bandwidth is not a closed problem, we recommend using the bandwidth obtained from cross-validation just as an initial value from which to try another. Then, the knowledge of the problem may be a criterium in choosing a suitable bandwidth.

The proposed method was applied to study the spatial distribution of soil pollutants in an industrialized area. Both statistical tests show a similar behavior, indicating the existence of spatial heterogeneity when the bandwidth is less than 30% of the data. Also the percentage of variance explained for the first four GWPCs is considerably higher than that explained for the first four global PCs in some samples. These results seem to confirm the suitability of implementing GWPCA instead of standard PCA.

The analysis of the winning variables for the first four geographically weighted compo-

nents proved to be very useful, since it allowed us to draw some interesting conclusions regarding the relationships between pollutants and their possible sources that would not be possible with a standard global principal components analysis.

Acknowledgments

The authors would like to express their gratitude for support received through the grants FC-15-GRUPIN14-033 of the Principado de Asturias (Spain) and MTM2014-55699-P of the Spanish Ministry of Science (both with FEDER support).

References

Adriano, D.C., 2003. *Trace Elements in Terrestrial Environments: Biogeochemistry, Bioavailability and Risks of Metals*. Springer, New York, NY, USA, 2nd edition.

Berciano, F.A., Domínguez, J. and Alvarez F.V., 1989. Influence of air pollution on extrinsic childhood asthma. *Annals of Allergy Asthma Immunology*, 62, 135-141.

Brunsdon, C., Fotheringham, A.S., and Charlton, M.E., 1998. Geographically weighted regression modelling spatial non-stationarity. The Statistician, 47(3), 431-443.

Burton, E.D., Phillips, I.R., and Hawker, D.W., 2005. Geochemical partitioning of copper, lead, and zinc in benthic, estuarine sediment profiles. *Journal of Environmental Quality*, 34, 263?273.

Charlton, M. Brunsdon, C., Demsar, U., Harris, P., Fotheringham, A.S., 2011. Principal Components Analysis: from Global to Local. 13th AGILE International Conference on Geographic Information Science 2010, Guimaraes, Portugal.

Demsar, U., Harris, P., Brunsdon, Ch., Fotheringham, A., and McLoone A.,2013. Principal Component Analysis on Spatial Data: An Overview. *Annals of the Association of American Geographers*,103 (1), 106-128.

Flor-Blanco, G., Flor, G., and Pando, L., 2013. Evolution of the Salinas-El Espartal and Xagó beach/dune systems in north-western Spain over recent decades: evidence for responses to natural processes and anthropogenic interventions. *Geo-Marine Letters*, 33(2-3), 143-157.

Fotheringham, A.S., Brunsdon, C., and Charlton, M.E., 2002. *Geographically weighted regression: the analysis of spatially varying relationships*. Wiley, Chichester, UK.

Gollini, I., Lu, B., Charlton, M., Brunsdon, Ch., and Harris, P., 2015. GWmodel: an R Package for Exploring Spatial Heterogeneity using Geographically Weighted Models 2015. *Journal of Statistical Software*, 63(17), 1-50.

Gallego, J.R. Ordóñez, A., and Loredo J., 2002. Investigation of trace element sources from an industrialized area (Avilés, northern Spain) using multivariate statistical methods. *Environmental International*, 27, 589-596.

Härdle W. and Bowman A.W., 1998. Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands. Journal of the American Statistical Association, 83(401), 102-110.

Harris P., Brunsdon C., and Charlton M., 2011. Geographically weighted principal component analysis. *International Journal of Geographical Information Science*, 25(10), 1717-1736.

Hope, A. ,2002. A Simplified Monte Carlo Significance Test Procedure. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(3), 582-598.

Ihaka, R., and Gentleman. R., 1996. R: A language for data analysis and graphics. itshape Journal of Computational and Graphical Statistics, 5, 299-314.

Jolliffe, I.T., 2002. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag New York.

Kumar S., Lal R., and Lloyd Ch.D., 2012. Assessing spatial variability in soil characteristics with geographically weighted principal components analysis. *Computational Geosciences*, 16(3), 827-835.

Lloyd, Ch.D., 2010. Analysing population characteristics using geographically weighted principal components analysis: a case study of Northern Ireland in 2001. *Computers, Environment and Urban Systems*, 34(5), 389-399.

Li, Y. Wang, Y.B., Gou, X., Su, Y.B., and Wang, G., 2006. Risk assessment of heavy metals in soils and vegetables around non-ferrous metals mining and smelting sites, Baiyin, China. *Journal of Environmental Sciences*, 6, 1124-1134.

Mattielli, N. Petit, J.C.J. Deboudt, K. Flament, P. Perdrix, E. Taillez, A., and Rimetz-Planchon, J. Zn isotope study of atmospheric emissions and dry depositions within a 5 km radius of a PbZn refinery. *Atmospheric Environment*, 43, 1265-1272.

Olade, M. A., 1987. *Dispersion of cadmium, lead and zinc in soils and sediments of a humid tropical ecosystem in Nigeria. Lead, mercury, cadmium and arsenic in the environment*. Edited by T. C. Hutchinson and K. M. Meema. John Wiley & Sons, 303-313.

R Core Team 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Tipping,M.E., and Bishop, Ch.M., 1999. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B*, 61(3), 611-622.

Nonparametric Regression: Optimal Local Bandwidth Choice. Journal of the Royal Statistical Society: Series B, 53(2), 453-464.

Wand, M.P. and Jones, M.C. 1995. *Kernel Smoothing*. Champan & Hall.

Weis, D. 2009. Zn isotope study of atmospheric emissions and dry depositions within a 5 km radius of a Pb-Zn refinery. *Atmospheric Environment*, 43, 1265-1272.

Figure Captions

Figure 1. Percentage of rejections for test statistics T_1 and T_2 on increasing a for nominal level of 5%, number of covariates $p = 5$ and $p = 10$(left and right panel, respectively).

Figure 2. Percentage of rejections for the test based on statistic T_2 for increasing values of a, nominal level 5% and number of covariates $p = 5$ and $p = 10$(left and right panel, respectively). Upper panel: rejections for sample size $n = 100$. Lower panel: rejections for sample size $n = 1000$.

Figure 3. Percentage of rejections for statistic test T_2 on increasing h for nominal level of 5% and number of covariates $p = 5$ and $p = 10$ (left and right panel, respectively). Upper panel: rejections for $a = 0.1$. Lower panel: rejections for $a = 0.2$.

Figure 4. Location of the samples overlaid on an orthophotography of the study area (Coordinate system: ETRS1989 UTM Zone 30N).

Figure 5. Cross-validation function as a function of h with four retained components.

Figure 6. Percentage of local variance for first four local components.

Figure 7. Winning variables for the first four GWPCA.