UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

ISABEL SOARES DINIZ DE OLIVEIRA

Visual Tools to Identify Influential Observations in Spatial Data

Recife

2021

ISABEL SOARES DINIZ DE OLIVEIRA

Visual Tools to Identify Influential Observations in Spatial Data

Master thesis presented to the Post-Graduate Program in Statistics at the Federal University of Pernambuco as part of the necessary requirements for obtaining a Master's Degree in Statistics.

Concentration area:: Applied Statistics

Advisor: Prof. Dra. Fernanda De Bastiani

Recife

2021

Catalogação na fonte
Bibliotecária Nataly Soares Leite Moro, CRB4-1722

O48v Oliveira, Isabel Soares Diniz de

Visual tools to identify influential observations in spatial data / Isabel Soares Diniz de Oliveira. – 2021.

60 f.: il., fig., tab.

Orientador: Fernanda de Bastiani.

Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2021.

Inclui referências.

1. Estatística Aplicada. 2. Análise de dados funcionais. 3. Dados espaciais influentes. 4. Semivariograma. I. Bastiani, Fernanda de (orientador). II. Título.

310 CDD (23. ed.) UFPE- CCEN 2022 - 14

ISABEL SOARES DINIZ DE OLIVEIRA

Visual Tools to Identify Influential Observations in Spatial Data

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 28 de outubro de 2021.

BANCA EXAMINADORA

FERNANDA DE BASTIANI
(UFPE)

GETULIO JOSE AMORIM DO AMARAL
(UFPE)

MANUEL JESUS GALEA ROJAS
(PUC-CHILE)

To my family, with love.

ACKNOWLEDGEMENTS

I would like to first thank God for lighting my way and giving me strength along my way here.

I am also especially grateful to my advisor Prof. Dra. Fernanda De Bastiani for believing in me and encouraging me in front of all the problems that came up during the Covid-19 Pandemic, as well as sharing her knowledge and giving me the opportunity to learn methods current to enrich my master thesis.

Dr. Mark G. Genton for contributing and sharing innovative methods.

I would like to express my gratitude to the master's board for having accepted to be part of and contribute to the improvement of the master's thesis.

I am also very grateful to my family for supporting me in every way possible, especially to my parents Risalva S. de Araújo and Edvaldo D. de Oliveira, who stayed by my side when I needed it most, and my grandmother Zuleide D. de Oliveira who always guided me and he was always interested in the progress of my thesis.

I am very grateful to my boyfriend Jonathan Romualdo S. A. Pereira for supporting and helping me correct my master's thesis, and my friends Flávia Helena M Vasconcelos and Lousiana R M Quinino for giving me moral support and also sharing their knowledge for my academic growth.

I am very grateful to Dra. Patricia L. E. Ospina for encouraging me and sharing her knowledge during my undergraduate degree in Statistics.

And to the National Council for Scientific and Technological Development (CNPq) for the financial support provided during my Master's Degree.

ABSTRACT

We adapted the hair-plot, proposed by Genton and Ruiz-Gazen (2010), to identify and visualize influential observations in spatial data. Three graphic tools were created: the bihair-plot, the principal components hair-plot and functional hair-plot. The first tool depict trajectories of the values of a spatial semivariance estimator when adding a perturbation to each observation of a vector of spatial data observed considering two lags. The second describes trajectories of the principal components of a spatial semivariance estimator values for all lags when each observation of data is perturbed, making it possible to identify influential observations in spatial data containing as much information as possible from the data set. The third is obtained from the values of the trace-semivariogram estimator when the data receive a disturbance. The estimators considered in the study were the sample semivariogram for univariate case, sample cross-semivariogram for bivariate case and sample trace-semivariogram for functional data. Another method used to obtain the cross-semivariogram was Minimum Volume Ellipsoid, which is more sensitive to outliers. Based on this, we observed that it is not possible to detect influential observations. We defined the quadratic form of the estimators and the influence function, in order to understand their behavior and properties. Finally, we make an application with these tools in the pollution data for the univariate case, complementing the results shown in Genton and Ruiz-Gazen (2010), the meuse data from the sp package for the bivariate case and average temperatures from the geofd package for the functional case.

Keywords: cross-semivariogram; functional data analysis; influential spatial data; principal components; semivariogram; trace-semivariogram.

RESUMO

Adaptamos o *hair-plot*, proposto por Genton and Ruiz-Gazen (2010), para identificar e visualizar observações influentes em dados espaciais. Três ferramentas gráficas foram criadas: o bihair-plot, os principais componentes do *hair-plot* e o *hair-plot* funcional. A primeira ferramenta descreve trajetórias dos valores de um estimador de semivariância espacial ao adicionar uma perturbação a cada observação de um vetor de dados espaciais observado considerando dois *lags*. O segundo descreve as trajetórias dos componentes principais de um estimador de semivariância espacial para todos os *lags* quando cada observação de dados é perturbada, tornando possível identificar observações influentes em dados espaciais contendo o máximo de informações possível do conjunto de dados. O terceiro é obtido a partir dos valores do estimador do *trace-semivariogram* quando os dados recebem uma perturbação. Os estimadores considerados no estudo foram o semivariograma de amostra para caso univariado, semivariograma cruzado de amostra para caso bivariado e *trace-semivariograma* amostral para dados funcionais. Outro método utilizado para obter o semivariograma cruzado foi o Elipsóide de Volume Mínimo, que é mais sensível a outliers. Com base nisso, observamos que não é possível detectar observações influentes. Definimos a forma quadrática dos estimadores e a função de influência, a fim de compreender seu comportamento e propriedades. Finalmente, fazemos uma aplicação com essas ferramentas nos dados de poluição para o caso univariado, complementando os resultados mostrados em Genton and Ruiz-Gazen (2010), os dados meuse do pacote sp para o caso bivariado e dados de temperaturas médias do pacote geofd para o caso funcional, inicialmente obtidas do Serviço Meteorológico do Canadá.

Palavras-chave: análise de dados funcionais; componentes principais; dados espaciais influentes; semivariograma; semivariograma cruzado; trace-semivariograma.

LIST OF FIGURES

Figura 1 – Distribution of reflectance values by quartile interval - pollution data. . . . 36

Figura 2 – Semivariogram amplitude graph for perturbed data represented by the bars, where the minimum and the maximum is the less and highest value of $\widehat{\gamma}(h)$ for $\zeta = [-40, 40]$, respectively, and semivariogram for original data represented by the points where $\zeta = 0$ - pollution data 37

Figura 3 – Hair-plots in the principal diagonal and the bihair-plot on top of the sample semivariogram on the reflectance residual values for spatial lag distance $h = 1, 2, 3, 4.$ and $\zeta \in (-40, 40)$, for the pollution data 38

Figura 4 – From left to right: PC hair-plots for PC1 and PC2 with 90.09% and 4.83% of sample semivariogram variability, respectively, and PC bihair-plot crossing PC1 and PC2 containing 94.92% of sample semivariogram cumulative variability, considering spatial lag distance $h = 1, 2, 3, 4.$ and $\zeta \in (-1, 1)$ - pollution data 39

Figura 5 – From left to right: PC hair-plots for PC1 and PC2 with 87.88% and 5.75% of sample semivariogram variability, respectively, and PC bihair-plot crossing PC1 and PC2 containing 93.63% of sample semivariogram cumulative variability, considering spatial lag distance $h = 1, \ldots, 8$ and $\zeta \in (-1, 1)$ - pollution data 39

Figura 6 – Hair-plots of the MVE sample semivariogram for the reflectance residual for $h = 1, 2, 3, 4$ - pollution data 40

Figura 7 – Distribution of reflectance values by quartile interval. The red point on the grid indicates the most influential point, corresponding to the maximum data value (observation $\#17$) - pollution data. 40

Figura 8 – Distribution map of zinc and lead values by quartile interval from left to right, respectively - meuse data. 41

Figura 9 – Estimates of method-of-moments semivariogram for (a) *log* of zinc and (b) *log* of lead, (c) original sample cross-semivariogram for *log* of zinc and *log* of lead represented by the points and cross-semivariogram amplitude graph of for *log* of zinc and *log* of lead perturbed, where the minimum and the maximum is the less and highest value of $\widehat{\gamma}(h)$ for $\zeta = [-1, 1]$, respectively, represented by the bars - meuse data 42

Figura 10 – Hair-plots in the principal diagonal and the bihair-plot on top diagonal of the sample cross-semivariogram on the values of zinc and lead, for spatial lag distance $h = 1, 2, 3, 4.$ and $\zeta \in (-1, 1)$. Blue values in the bihair-plot indicate the curve for which $\zeta < 0.$ - meuse data 43

Figura 11 – From left to right: PC hair-plots for PC1 and PC2 with $65, 96\%$ and $20, 89\%$ of the sample cross-semivariogram variability, respectively, and PC bihair-plot crossing PC1 and PC2 containing 85.85% of the sample cross-semivariogram cumulative variability, considering lags distance $h = 1, 2, 3, 4.$ and $\zeta \in (-1, 1)$ - meuse data 45

Figura 12 – From left to right: PC hair-plots for PC1 and PC2 with 67.83% and 11.21% of the sample cross-semivariogram variability, respectively, and PC bihair-plot crossing PC1 and PC2 containing 79.04% of the sample cross-semivariogram cumulative variability, considering lags distance $h = \{1, \ldots, 12\}$ and $\zeta \in (-1, 1)$ - meuse data 45

Figura 13 – Hair-plots of the MVE sample cross-semivariogram for the $\log$ of zinc and $\log$ of lead for $h = 1, 2, 3, 4$ - meuse data 45

Figura 14 – Distribution map of zinc and lead values by quartile interval from left to right, respectively. The points in red correspond to the most influential observation ($\#54$) and a possible influential point ($\#76$) ($\zeta > 0$), at locations $(179973, 332255)$ and $(179095, 330636)$ given in meters. - meuse data. . . 46

Figura 15 – Averages location of daily temperature curves observed at 35 weather stations of the Canadian Maritime provinces - maritimes data. 48

Figura 16 – Matplot of daily temperature curves observed at 35 weather stations of the Canadian Maritime provinces - maritimes data 48

Figura 17 – Smoothed data of daily temperature curves obtained by B-spline with 65 basis functions - maritimes data. 49

Figura 18 – Trace-semivariogram amplitude graph for perturbed data represented by the bars, where the minimum and the maximum is the less and highest value of $\widehat{\gamma}(h)$ for $\zeta = [-1, 1]$, respectively - maritimes data 49

Figura 19 – Hair-plots in the principal diagonal and the bihair-plot on top diagonal of the sample trace-semivariogram on the values of maritimes data, for spatial lag distance $h = 1, 2, 3, 4.$ and $\zeta \in (-1, 1)$. The black curve represents the observation that was influential by analyzing a lag among those taking part in the bihairplot - maritimes data 50

Figura 20 – From left to right: PC hair-plots for PC1 and PC2 with 79.52% and 16.90% of sample trace-semivariogram variability, respectively, and PC bihair-plot crossing PC1 and PC2 containing 96.43% of sample trace-semivariogram cumulative variability, considering spatial lag distance $h = \{1, 2, 3, 4\}$ and $\zeta \in [-1, 1]$ - maritimes data 51

Figura 21 – From left to right: PC hair-plots for PC1 and PC2 with 57.88% and 26.72% of sample trace-semivariogram variability, respectively, and PC bihair-plot crossing PC1 and PC2 containing 84.68% of sample trace-semivariogram cumulative variability, considering spatial lag distance $h = 1, \ldots, 4$ and $\zeta \in (-1, 1)$ - maritimes data 51

Figura 22 – From left to right: matplot highlighting in red the points corresponding to the observed location 22, 29 and 16, respectively- maritimes data 52

Figura 23 – Averages locations of daily temperature at 35 weather stations. The points in red correspond to possible influential points (#16,#22 and #29), at locations $(-64.85, 44.23)$, $(-63.5, 44.63)$ and $(-66.47, 45.85)$, respectively - maritimes data. 52

LIST OF TABLES

Tabela 1 – Descriptive statistics values of pollution data. 36
Tabela 2 – Descritive statistics of method-of-moments semivariogram and the observation corresponding to the maximum value of the estimates for each lag ($\zeta > 0$). 36
Tabela 3 – Descritive statistics of method-of-moments cross-semivariogram and the observation corresponding to the maximum value of the estimates for each lag ($\zeta > 0$) - meuse data. 43
Tabela 4 – Descritive statistics of method-of-moments trace-semivariogram and the observation corresponding to the maximum value of the estimates for each lag ($\zeta > 0$) - pollution data. 47

LIST OF ACRONYMS AND ABBREVIATIONS

FDA	Functional Data Analysis
MVE	Minimum Volume Ellipsoide
PC	Principal Component
PCA	Principal Component Analysis

CONTENTS

1 INTRODUCTION . . . 15

1.1 MOTIVATION . . . 16

1.2 CONTRIBUTIONS . . . 17

1.3 FUNCTIONAL DATA . . . 18

1.4 SPATIAL DATA . . . 19

2 SPATIAL VARIABILITY ESTIMATORS ON SAMPLE POINTS AND FUNCTIONAL DATA . . . 21

2.1 SEMIVARIOGRAM . . . 21

2.1.1 Method-of-moments semivariogram . . . 21

2.1.2 Method-of-moments semivariogram on quadratic form . . . 22

2.2 CROSS-SEMIVARIOGRAM . . . 22

2.2.1 Method-of-moments cross-semivariogram . . . 23

2.2.2 Method-of-moments cross-semivariogram on quadratic form . . . 23

2.2.3 Minimum Volume Ellipsoid cross-semivariogram . . . 24

2.3 METHOD-OF-MOMENTS TRACE-SEMIVARIOGRAM . . . 24

3 PRINCIPAL COMPONENTS ANALYSIS . . . 26

3.1 PRINCIPAL COMPONENTS . . . 26

3.2 COVARIANCE AND CORRELATION MATRIX . . . 27

3.3 MINIMUM NUMBER OF PRINCIPAL COMPONENTS . . . 28

4 INFLUENTIAL ANALYSIS . . . 29

4.1 EMPIRICAL, LOCAL AND ASYMPTOTIC INFLUENCE . . . 29

4.2 INFLUENCE ON QUADRATIC FORM . . . 30

4.3 GRAPHICAL TOOLS TO VISUALIZE INFLUENTIAL SPATIAL DATA . . . 31

4.3.1 Bihair-plot . . . 31

4.3.2 Principal components hair-plot . . . 31

5 APPLICATIONS . . . 35

5.1 POLLUTION DATA . . . 35

5.2 MEUSE DATA . . . 41

5.3 MARITIMES DATA . . . 47

6 SUPPLEMENTARY MATERIALS . . . 53

7 CONCLUSIONS 54

8 FUTURE WORKS 55

9 CONCLUSIONS 56

10 FUTURE WORKS 57

REFERENCES 58

1 INTRODUCTION

Spatial statistics is an area that studies data considering the space in which it was sampled in such a way that the observation is associated with its location. Spatial dependence is detected using spatial statistical techniques that identify patterns based on data distribution and spatial variability. According to Cressie (1993), the closer the data the more similar they are, that is, the variability increases as the distance between the data increases. Several areas of knowledge use geostatistics to identify and describe phenomena that behave according to their location, such as: soil science, geology, forestry, agriculture, epidemiology, etc. To model the spatial distribution of the Covid-19 infection risk and assuming that the uncertainty of spatial predictions is rarely studied, Azevedo et al. (2020) proposed to apply a direct block sequential simulation. They add that the spatial analysis of the phenomenon's dynamics over time can be studied from the slope of the linear regression line in the short term, or from the functional analysis of the data. Therefore, it was possible to identify areas with a higher risk of infection by providing local estimates of the probabilities such that they exceed the threshold obtained from the simulated scenarios.

Martín, Arias and Corbí (2006) applied multivariate geostatistical methods to identify locations with higher concentrations of heavy metals in agricultural topsoils from geostatistical models and to estimate concentrations on non-sampled locations. They concluded that heavy metal concentrations are not high enough to be pollutants in spite of anthropic activity, but they found local anomalies of some heavy metals associated with anthropogenic activities. Cortés-D, Camacho-Tamayo and Giraldo (2016) used functional geostatistics models to predict the resistance of soil penetration, in which non-parametric smoothing functions were fitted to the data. They noticed that the behavior of the observed and predicted data are similar, and that the model fit becomes better and more homogeneous as depth increases.

Identifying influential points and *outliers* in spatial data is one of the important steps in exploratory spatial analysis and diagnostic analysis, as such observation can change the results obtained through kriging or coktiging and change the structure that describes spatial dependence. An observation is influential whenever a change in its value radically changes the estimate or some property of the fitted model, and a *spatial outlier* when the observation is extreme in relation to its neighbors (JONATHAN et al., 2016). An outlier can be influential, such that an extreme observation can influence a model's estimates. On the other hand, an influential point may not be an outlier, which can be identified in the diagnostic analysis of the applied model. It is common to find studies in which data is perturbed excluding observations, such as in the context of linear regression models diagnostic and methods for generalized least squares estimators that included Cook's distance, assuming the known and fixed covariance matrix for a scalar multiplier (COOK, 1977; MARTIN, 1992). Fox (1972) generalized Cook's method for the dependent data in time series, introducing two types of outliers.

In order to assess the sensitivity of the maximum likelihood estimator in elliptical spatial linear models due to small contamination, De Bastiani et al. (2015) used of the local influence and concluded that outliers strongly influence the spatial dependence structure. With the same objective of detecting observations that influence the values obtained from the maximum likelihood estimator of linear spatial Gaussian models, Borssoi et al. (2011) evaluates the influence perturbing the matrix of exploratory variables. Baba et al. (2021) adapted some classical methods making them more robust for detecting influential points in spatial regression models and compared them with Cook's distance, showing the advantage of such methods for these types of models.

The estimators considered in this master's thesis were based on the moments-of-methods semivariogram, proposed by Matheron (1963), methods-of-moments cross-semivariogram (LARK, 2003), and the moments-of-methods trace-semivariogram (GIRALDO; MATEU; DELICADO, 2012) for the univariate, bivariate and functional case, respectively. These estimators are more sensitive to outliers, and it is expected that changes in input values due to an additive perturbation will change the estimate more drastically compared to estimators that are more robust to outliers. As an example, a highly robust cross-semivariogram estimator proposed by Lark (2003) was used, whose structure contains the covariance matrix obtained by the Minimum Volume Ellipsoid (MVE) (see Aelst and Rousseeuw (2009)).

1.1 MOTIVATION

When the study involves dependent observations, as in the case of spatial data, the influence function includes the joint distribution of the data, therefore the it was evaluated from a method involving additive perturbation. Genton and Ruiz-Gazen (2010) proposed a tool to visualize influential observations in the context of dependent data based on the study of the data perturbation effect on the estimators $\widehat{\theta}(\mathbf{Z})$ of a parameter $\theta(\mathbf{Z})$, and they introduced the *hair-plot* in order to detect and analyse influential points. For this tool development, they defined an empirical influence based on additive perturbation $\mathbf{Z}[i, \zeta], i = 1, \ldots, n$, in the context of dependent data, considering a perturbation value $\zeta \in \mathbb{R}$, that allowed to obtain more information about behavior of estimators $\widehat{\theta}(\mathbf{Z}[i, \zeta])$. In other words, they describe all trajectories of the $\widehat{\theta}(\cdot)$ values by adding a perturbation to each observation of the data. They also proposed two influential measures: local and asymptotic influece of i-th observation, such that a largest absolute value of first measure indicates the most influential observation and the second indicates the influence on the estimator value when ζ is large.

In this present master's thesis, $\mathbf{Z}(\mathbf{s})$ is a spatial process and $\widehat{\theta}(\cdot)$ is the method-of-moments semivariogram (MATHERON, 1963) for univariate spatial data point, the method-of-moments and MVE cross-semivariogram (LARK, 2003) for bivariate spatial data point and method-of-moments trace-semivariogram for functional data (GIRALDO; MATEU; DELICADO, 2012). They measure the degree of spatial dependence evaluated in a vector $\mathbf{h}$ of distances or *lags*. The term *lag* is used when referring to a sequence of the number of breaks in the distance interval, such that $h = 1$ indicates the minimum distance between the points that contains 30 pairs, and the maximum h represents the cutoff of up to 50% of the maximum distance between the points. The MVE cross-semivariogram is more robust to influential observations, such that its estimate is not affected by those observations (see Lark (2003)).

Our goal is to adapt the visual tool *hair-plot* to find influential observations taking information from $\widehat{\theta}(\mathbf{Z}[i, \zeta])$ values estimates through: *bihair-plot*, considering estimate of two lags $\mathbf{h}$ together; principal components hair-plot, analysing all values of $\mathbf{h}$; and hair-plot functional perturbed all values in one site. Bihair-plot, based on hair-plot, checks if there is an influential observation analysing the lags $\mathbf{h}$ pairwise. In this way, this graphical tool can be seen as a function of empirical influence evaluated at two distances. in order to identify an influential observation considering spatial factor, we analyze $\widehat{\theta}(\cdot)$ values for all values of $\mathbf{h}$ together in hair-plot, through principal components analysis (PCA).

1.2 CONTRIBUTIONS

The contributions of this master thesis are to adapt the hair-plot to:

- Identify influential points by considering two paired lags from bihair-plot;
- Define a form for the perturbation function based on principal component analysis so that it is possible to load as much information from the estimates considering all lags

and identify the most influential point from principal components hair-plot;

- Add perturbation to functional data by identifying a location with an influential function from functional hair-plot.

1.3 FUNCTIONAL DATA

Ramsay and Dalzell (1991) introduced the concept of functional data, which consists of observations represented by functions, that is, the *i-th* observation of the data set is expressed by a real function, $Z(t), i = 1, \ldots, n, t \in T$, where $T \subseteq \mathbb{R}$, and n corresponds to the number of observations. A multivariate data analysis is not appropriate when each curve is observed separately, and the data is smoothed to study its behavior, treating it this way as a continuous functions defined in a common interval.

Functional data analysis (FDA) can be applied for the following reasons: functional data is increasingly common in applied contexts, and smoothing and interpolation methods can produce functional results from a set of observations; there are some problems that become easier to interpret when dealing with data in a functional way when, for example, data is used to estimate a function or its derivatives; when it is necessary to smooth out multivariate data that results in functional processes, that is, we can describe a set of observation with a function. Ramsay and Silverman (2005) describes several methods and techniques for handling functional data and reports that the main objectives of the FDA are: identify patterns in the data; study the dynamics of the data; highlight expressive aspects; represent the data to facilitate future analysis; explain the behavior of the output variables (dependent variable) using information from the input variable (independent variable).

The FDA can be divided into three parts: exploratory, confirmatory and predictive. According to Genton and Sun (2020), in exploratory FDA, the following visualization tools can be used for univariate and multivariate data analysis, respectively: the functional boxplots and surface boxplots; magnitude-shape plots, two-stage functional boxplots and trajectory functional boxplots.

According to Ferraty and Vieu (2006), a functional variable is defined as a random variable in an infinite dimensional space, also called a functional space. Let $z_1, ..., z_n$ be observed values of $Z_1, ..., Z_n$, n identically distributed functional variables of Z and let $T = [t_{min}, t_{max}] \subseteq \mathbb{R}$. Functional data are elements of:

$$L_2(T) = \{Z : T \to \mathbb{R}, \text{where} \int_T Z(t)^2 dt < \infty\},$$

such that $L_2(T)$ can be rewritten as the inner product defined an Euclidian space: $\langle x, y\rangle = \int_T x(t)y(t)dt$.

Let a functional variable $Z_s(t)$ for all $s \in D$. A functional random process is defined as $\{Z_s(t) : s \in D \subseteq \mathbb{R}, t \in T \subseteq \mathbb{R}\}$. When data is generated from a large number of measurements (over space, for example), each observation can be expressed from a non-parametric function observed in terms of K basis functions (GIRALDO; MATEU; DELICADO, 2012):

$$z_s(t) = \sum_{l=1}^{K} a_{il}\phi_l(t) = \mathbf{a}_i^\top \boldsymbol{\phi}(t), \tag{1.1}$$

where: $i = 1 \ldots n, a_i = (a_{i1}, \ldots, a_{iK})$ and $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots, \phi_K(t)), \boldsymbol{\phi}(t)$ represent basis functions.

The expression 1.1 represents the truncated versions of Fourier series (periodic data) or B-splines expansions (non-periodic data).

1.4 SPATIAL DATA

It is often interesting to understand a phenomenon taking into account the space in which it behaves, such as the spread of a disease in a certain region (AZEVEDO et al., 2020) or the concentration of metals that affect a agricultural topsoil (MARTÍN; ARIAS; CORBÍ, 2006). When data is spatially referenced, it is important to study its behavior in the space for decision making where the phenomenon behaves atypically. Thus, it is important to apply methods that identify spatial patterns and carry information from the spatial correlation structure

In geostatistics, a spatial random field $\{Z(\mathrm{s}) : \mathrm{s} \in D_{\mathrm{s}} \subset \mathbb{R}^d\}$ is defined as a stochastic process, where D is a subset of the d-dimensional Euclidean space. It can be assumed that the covariance between two random variables depends on the spatial lag distance $\mathbf{h}$ between their locations. It is common to assume stationarity in this type of process, and three types of processes are defined here, according to distribution, expectation, and variance (CASTRO, 2013):

- **Strictly stationary**: $(Z_{\mathbf{s}_1}, Z_{\mathbf{s}_2}, \ldots, Z_{\mathbf{s}_k})$ has the same joint distribution as $(Z_{\mathbf{s}_1+\mathbf{h}}, Z_{\mathbf{s}_2+\mathbf{h}},$

$\ldots, Z_{\mathbf{s}_k+\mathbf{h}})$ for locations $\{\mathbf{s}_1, ..., \mathbf{s}_k\}$:

$$F_{(X_{s_1}, X_{s_2}, ..., X_{s_k})}(x_{s_1}, x_{s_2}, \ldots, x_{s_k}) = F_{(X_{s_1+\mathbf{h}} X_{s_2+\mathbf{h}}, ..., X_{s_k+\mathbf{h}})}(x_{s_1+\mathbf{h}}, x_{s_2+\mathbf{h}}, \ldots, x_{s_k+\mathbf{h}});$$

- **Weakly stationary**: E$(Z_\mathbf{s}) = \mu$ and Cov$(Z_\mathbf{s}, Z_{\mathbf{s}+h}) = C(h)$, where $C(h)$ can be called a covariogram;
- **Intrinsically stationary**: assume E$(Z_{\mathbf{s}+h} - Z_\mathbf{s}) = 0$ and get Var$(Z_{\mathbf{s}+h} - Z_s) =$ E$(Z_{\mathbf{s}+h} - Z_\mathbf{s})^2 = 2\gamma(h)$, where $\gamma(h)$ is called a semi-variance function.

Still according to Castro (2013), they have the following relationship:

strictly stationary $\Rightarrow$ weakly stationary $\Rightarrow$ stochastic processes

But the return is not necessarily true.

2 SPATIAL VARIABILITY ESTIMATORS ON SAMPLE POINTS AND FUNCTIONAL DATA

2.1 SEMIVARIOGRAM

The *semivariogram*, measures the level of dependence between two samples separated by distance vector $\mathbf{h} \in \mathbb{R}^d$, with sample locations $(\mathbf{s})$ and $(\mathbf{s}+\mathbf{h})$, associated with a regionalized variable $Z(\mathbf{s})$, where $\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ is a intrisically stacionary process defined on a domain D. According to Cressie (1989), the semivariogram can be defined as:

$$\gamma(\mathbf{h}) = \frac{1}{2}Var[Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})] = \frac{1}{2}E[Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})]^2, \quad \forall \mathbf{s}, \mathbf{s}+\mathbf{h} \in D \tag{2.1}$$

The $2\gamma(\mathbf{h})$ is referred to as a *variogram*.

There are several ways to estimate the semivariogram. One of them is the method of moments, introduced by Matheron (1963), commonly used in the literature. However, this estimator is not robust to outliers. Therefore, other estimators were proposed in order to be more robust to outliers, such as the one by Cressie and Hawkins (1980), the variogram fitting by generalized least squares of Genton (1998), and the pairwaise relative variogram described by Bai and Deutsch (2020).

In this master's thesis, we applied the method-of-moments semivariogram in order to identify influential points. Finally, the spatial dependence is analyzed from the graph of semivariogram estimates *versus* distance h.

2.1.1 Method-of-moments semivariogram

The method-of-moments semivariogram, proposed by Matheron (1963) is given by:

$$\widehat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}[Z(\mathbf{s}_i+\mathbf{h}) - Z(\mathbf{s}_i)]^2, \tag{2.2}$$

where

- $\widehat{\gamma}(\mathbf{h})$ is the value of the semivariogram estimate;
- $Z(\mathbf{s}_i)$ is the value of the variable Z in position $\mathbf{s}_i$;
- $Z(\mathbf{s}_i+\mathbf{h})$ is the value of the variable Z in position $\mathbf{s}_i+\mathbf{h}$;

- $N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{h}\}$ is the number of pairs separated by a given distance $\mathbf{h}$.

For an irregular sampled data grid, $N(\mathbf{h})$ can be write as $\{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j \in T(\mathbf{h})\}$, where $T(\mathbf{h}) \subset R^d$ surrounded $\mathbf{h}$ (CRESSIE, 1989).

2.1.2 Method-of-moments semivariogram on quadratic form

For a better understanding of the structure and properties of the estimators, Genton (1998) restructured the Equation 2.2 and proposed the following theorem in order to define the expected value, the variance, the covariance and the correlation of the semivariogram. Let $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^\top$ a spatial data vector and $A(\mathbf{h})$ a spatial design matrix, the quadratic form of semivariogram estimator by method-of-moments (Equation 2.2) is:

$$\widehat{\gamma}(\mathbf{h})_Q = \frac{1}{2}\mathbf{Z}^\top A(\mathbf{h})\mathbf{Z} \tag{2.3}$$

Let $\mathbf{z}$ a random vector, $\mathbb{E}(\mathbf{z}) = \mu \mathbb{1}_n$ and $Var(\mathbf{z}) = \mathbf{\Sigma}$. Then:

(a) $\mathbb{E}(\widehat{\gamma}(\mathbf{h})) = \frac{1}{2}\text{tr}[A(\mathbf{h})\mathbf{\Sigma}]$;

if $\mathbf{Z}$ is also Gaussian, then:

(b) Var$(\widehat{\gamma}(\mathbf{h})) = \text{tr}[A(\mathbf{h})\mathbf{\Sigma}A(\mathbf{h})\mathbf{\Sigma}]$;

(c) Cov$(\widehat{\gamma}(\mathbf{h}_1), \widehat{\gamma}(\mathbf{h}_2)) = \text{tr}[A(\mathbf{h}_1)\mathbf{\Sigma}A(\mathbf{h}_2)\mathbf{\Sigma}]$;

(d) Corr$(\widehat{\gamma}(\mathbf{h}_1), \widehat{\gamma}(\mathbf{h}_2)) = \frac{tr[A(\mathbf{h}_1)\mathbf{\Sigma}A(\mathbf{h}_2)\mathbf{\Sigma}]}{2\sqrt{tr[A(\mathbf{h}_1)\mathbf{\Sigma}A(\mathbf{h}_1)\mathbf{\Sigma}]tr[A(\mathbf{h}_2)\mathbf{\Sigma}A(\mathbf{h}_2)\mathbf{\Sigma}]}}$

where $A(\mathbf{h})$ can be composed by superposing identity matrices I_{n-h}, as below:

$$A(h) = 1/(n-h)\begin{pmatrix} I_{n-h} & -I_{n-h} \\ -I_{n-h} & I_{n-h} \end{pmatrix}$$

The *Proof* of **Theorem** 2.1.2 is described in detail by Genton (1998).

2.2 CROSS-SEMIVARIOGRAM

The *Cross-semivariogram* represents an association between two regionalized variables Z_u and Z_v, which $\{Z_u(\mathbf{s}), Z_v(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ are a intrisically stacionary process defined on a domain D, and it measures the association between them. The cross-semivariogram is defined as:

$$\gamma(\mathbf{h}) = \frac{1}{2}E[Z_u(\mathbf{s}+\mathbf{h}) - Z_u(\mathbf{s})][Z_v(\mathbf{s}+\mathbf{h}) - Z_v(\mathbf{s})] \tag{2.4}$$

where $(\mathbf{s})$ and $(\mathbf{s}+\mathbf{h})$ are sample locations and h represents the distance between the sample locations (see Lark (2003)).

2.2.1 Method-of-moments cross-semivariogram

A cross-semivariogram estimated from spatial data vectors, $\mathbf{Z}_u = (Z_u(\mathbf{s}_1), \ldots, Z_u(\mathbf{s}_n))^\top$ and $\mathbf{Z}_v = (Z_v(\mathbf{s}_1), \ldots, Z_v(\mathbf{s}_n))^\top$, by the method-of-moments can be defined as Lark (2003):

$$\widehat{\gamma}_{u,v}(\mathbf{h}) = \widehat{\gamma}_{v,u}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}[Z_u(s_i+\mathbf{h}) - Z_u(s_i)][Z_v(s_i+\mathbf{h}) - Z_v(s_i)] \tag{2.5}$$

- $\widehat{\gamma}_{u,v}(\mathbf{h})$ is the value of the cross-semivariogram estimate;
- $Z_u(\mathbf{s}_i)$ and $Z_v(\mathbf{s}_i)$ are the values of the variable Z_u and Z_v, respectively, in position $\mathbf{s}_i$;
- $Z_u(\mathbf{s}_i+\mathbf{h})$ and $Z_v(\mathbf{s}_i+\mathbf{h})$ is the value of the variable Z_u and Z_v, respectively, in position $\mathbf{s}_i+\mathbf{h}$;
- $N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{h}\}$ is the number of pairs separated by a given distance $\mathbf{h}$. For an irregular sampled data grid, $N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j \in T(\mathbf{h})\}$, where $T(\mathbf{h}) \subset \mathbb{R}^d$ surrounded $\mathbf{h}$ (CRESSIE, 1989).

In this case, when the reach and threshold are evaluated, we are interested in studying the maximum distance of spatial dependence and the approximation of the covariance between the two variables, respectively.

The Cauchy-Schwartz relation, given by: $|\gamma_{u,v}| = \sqrt{\gamma_u\gamma_v}$; all distances h considered in a co-kriging process must be guaranteed.

2.2.2 Method-of-moments cross-semivariogram on quadratic form

Let $A(\mathbf{h})$ a spatial design matrix, from the expression in Equation 2.5, we can obtain the cross-semivariogram estimated by method-of-moments quadratic form is:

$$\widehat{\gamma}_{\mathsf{uv}}(\mathbf{h})_Q = \frac{1}{2}\mathbf{Z}_{\mathsf{u}}^{\top} A(\mathbf{h})\mathbf{Z}_{\mathsf{v}} \tag{2.6}$$

So, from the Theorem 2.1.2, we can obtain the first and second moments as follows:

Let $\mathbf{Z}_u$ and $\mathbf{Z}_v$ random vectors, $\mathbb{E}(\mathbf{Z}_u) = \mu_u \mathbb{1}_n$ and $\mathbb{E}(\mathbf{Z}_v) = \mu_v \mathbb{1}_n$, and $\text{Cov}(\mathbf{Z}_u, \mathbf{Z}_v) = \mathbf{\Sigma}_{uv}$. Then:

(a) $\mathbb{E}(\widehat{\gamma}(\mathbf{h})) = \frac{1}{2}\ \text{tr}[A(\mathbf{h})\mathbf{\Sigma}_{uv}]$;

if $\mathbf{z}$ is also Gaussian, then:

(b) $\text{Var}(\widehat{\gamma}(\mathbf{h})) = \text{tr}[A(\mathbf{h})\mathbf{\Sigma}_{uv}A(\mathbf{h})\mathbf{\Sigma}_{uv}]$;

(c) $\text{Cov}(\widehat{\gamma}(\mathbf{h})) = \text{tr}[A(\mathbf{h}_1)\mathbf{\Sigma}_{uv}A(\mathbf{h}_2)\mathbf{\Sigma}_{uv}]$;

(d) $\text{Corr}(\widehat{\gamma}(\mathbf{h}_1), \widehat{\gamma}(\mathbf{h}_2)) = \frac{tr[A(\mathbf{h}_1)\mathbf{\Sigma}_{uv}A(\mathbf{h}_2)\mathbf{\Sigma}_{uv}]}{2\sqrt{tr[A(\mathbf{h}_1)\mathbf{\Sigma}A(\mathbf{h}_1)\mathbf{\Sigma}_{uv}]tr[A(\mathbf{h}_2)\mathbf{\Sigma}A(\mathbf{h}_2)\mathbf{\Sigma}_{uv}]}}$

2.2.3 Minimum Volume Ellipsoid cross-semivariogram

When the data set contains outliers, the semivariogram and cross-semivariogram obtained by method-of-moments can result in overestimation of error variance by cokriging. Lark (2003) proposed the MVE, a cross-variogram estimator as a function of the robust *p-variate* variance-covariance matrix $\mathbf{C}$ introduced by Rousseeuw (1984), and given by:

$$\widehat{\mathbf{\Gamma}}^{MVE}(\mathbf{h}) = \frac{1}{2}\widehat{C}_{MVE}\left[\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^{N(\mathbf{h})}\right]^{\top}, \tag{2.7}$$

where

- $\mathbf{y}^i(\mathbf{h}) = \{\mathsf{y}_1^i(\mathbf{h}), \mathsf{y}_2^i(\mathbf{h}), ..., \mathsf{y}_p^i(\mathbf{h})\}$
- $\mathsf{y}_u^i(\mathbf{h}) = Z_u(\mathbf{s}_i) - Z_u(\mathbf{s}_i + \mathbf{h})$ is a paired difference.

A cross-variogram estimate $\widehat{\gamma}_{u.v}^{MVE}(\mathbf{h})$ corresponds to the element $\{u, v\}$ of the matrix $\widehat{\mathbf{\Gamma}}^{MVE}(\mathbf{h})$.

2.3 METHOD-OF-MOMENTS TRACE-SEMIVARIOGRAM

Assuming that $m(t)$ over D the mean function of $Z_s(t)$, it is follow that (GIRALDO; MATEU; DELICADO, 2012):

$$\gamma(h) = \frac{1}{2}\mathbb{E}\left[\int_T (Z_{s_i}(t) - Z_{s_j}(t))^2 dt\right], \text{where : } s_i, s_j \in D, \text{ and } h = \|s_i - s_j\|. \tag{2.8}$$

Therefore, following the expression of the classic Mantheron semivariogram, the semivariogram estimator for the functional data is given by:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2|N(\mathbf{h})|} \sum_{i,j \in N(\mathbf{h})} \int_T (Z_{s_i}(t) - Z_{s_j}(t))^2 \tag{2.9}$$

where:

- $\widehat{\gamma}(\mathbf{h})$ is the value of the trace-semivariogram estimate;
- $Z_{\mathbf{s}_i}(t)$ is the value of the functional variable $Z(t)$ in position $\mathbf{s}_i$;
- $Z_{\mathbf{s}_i+\mathbf{h}}(t)$ is the value of the functional variable $Z(t)$, in position $\mathbf{s}_i + \mathbf{h}$;
- $N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{h}\}$ is the number of pairs separated by a given distance $\mathbf{h}$. For an irregular sampled data grid, $N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j \in T(\mathbf{h})\}$, where $T(\mathbf{h}) \subset \mathbb{R}^d$ surrounded $\mathbf{h}$ (CRESSIE, 1989).

3 PRINCIPAL COMPONENTS ANALYSIS

Increasingly, multivariate data sets are more present, and concomitantly difficulties arise to interpret and graph them. How to get around this problem? Principal Component Analysis (PCA) appears with the goal of facilitating the interpretation of any data, reducing its dimensionality through the principal component (PC), which are linear combinations of the correlated variables of the data set. The PCs are not correlated with each other, and are ordered in such a way that the former contains most of the data variation. They are found through the decomposition of the centralized data matrix, reducing to a linear optimization problem, in such a way that minimizes the size of the data, maximizing the variability subject to certain restrictions. These can represent the original data, As the use of PCA is descriptive in nature, the observed data can be used regardless of their distribution. The assumption of normality can be assumed for inferential purposes (JOLLIFFE; CADIMA, 2016; EVERITT; HOTHORN, 2011).

3.1 PRINCIPAL COMPONENTS

The PCA has as main objective to explain the maximum variation present in a data set with n observations and correlated numerical variables $\mathbf{x}_1, \ldots, \mathbf{x}_p$ through a new set of uncorrelated variables $\mathbf{y}_1, \ldots, \mathbf{y}_p$, such that each y_i, $i = 1, \ldots, p$, is a linear combination of q variables $(q \leq p)$. This new variable is called the principal component (PC) and ordered in such a way that the first component contains the maximum variance of the $\mathbf{X}_{n\times p}$ data matrix.

Thus, the first component $(\mathbf{y}_1)$, or PC1, has the following expression:

$$\mathbf{y}_1 = \beta_{11}\mathbf{x}_1 + \beta_{12}\mathbf{x}_2 + \ldots + \beta_{1p}\mathbf{x}_p = \mathbf{X}\boldsymbol{\beta}_1,$$

where $\boldsymbol{\beta}_1^\top\boldsymbol{\beta}_1 = 1$, Var$(\mathbf{y}_1) = \boldsymbol{\beta}_1^\top\mathbf{S}\boldsymbol{\beta}_1$ is the sample variance of $\mathbf{y}_1$, $\mathbf{S}$ is the $p \times p$ sample covariance matrix of $\mathbf{X}$.

Similarly, we have the expression for the second component $\mathbf{y_2}$, or PC2, given by:

$$\mathbf{y}_2 = a_{21}\mathbf{x}_1 + a_{22}\mathbf{x}_2 + \ldots + a_{2p}\mathbf{x}_p = \mathbf{X}\boldsymbol{\beta}_2,$$

subject to constraints: $\begin{cases} \boldsymbol{\beta}_2^\top\boldsymbol{\beta}_2 & = 1 \\ \boldsymbol{\beta}_2^\top\boldsymbol{\beta}_1 & = 0 \end{cases}$

And in the same way, the j-th PC is defined as $\mathbf{y}_j = \mathbf{X}\boldsymbol{\beta}_j$, subject to constraints:

$$\begin{cases} \boldsymbol{\beta}_j^\top \boldsymbol{\beta}_j = 1 \\ \\ \boldsymbol{\beta}_j^\top \boldsymbol{\beta}_l = 0 \quad (j < l), l = 1, \ldots, p. \end{cases}$$

where Var$(\mathbf{y}_j) = \boldsymbol{\beta}_j^\top \mathbf{S} \boldsymbol{\beta}_j$.

Then there is a problem of maximizing a function of multiple variables, subject to at least one constraint, therefore the method of *Lagrange multipliers* is used in such a way that $\boldsymbol{\beta}$ is the eigenvector, and $\mathbf{S}$ corresponding to this matrix's largest eigenvalue (see Everitt and Hothorn (2011)). Then, we want to maximizing $\boldsymbol{\beta}_j^\top \mathbf{S} \boldsymbol{\beta}_j - \lambda(\boldsymbol{\beta}_j^\top \boldsymbol{\beta}_j - 1)$, such that λ is a *Langrange multiplier*, finding λ from the equation:

$$\mathbf{S}\boldsymbol{\beta}_j = \lambda_j \boldsymbol{\beta}_j \Longleftrightarrow \mathbf{S}\boldsymbol{\beta}_j - \lambda_j \boldsymbol{\beta}_j = 0, \tag{3.1}$$

thus, the covariance between $\mathbf{X}\boldsymbol{\beta}_j$ and $\mathbf{X}\boldsymbol{\beta}_i$ is

$$\boldsymbol{\beta}_l^\top \mathbf{S} \boldsymbol{\beta}_j = \lambda_j \boldsymbol{\beta}_l^\top \boldsymbol{\beta}_j = \begin{cases} \lambda_j & l = j \\ \\ 0 & (l \neq j). \end{cases} \tag{3.2}$$

The eigenvectors $\boldsymbol{\beta}_j$ and the linear combinations $\mathbf{X}\boldsymbol{\beta}_j$ are called *PC loadings* and *PC scores*, respectively.

From the expression of Equation 4.6, the total variance of the q principal components is given by: $\sum_{j=1}^{p} \lambda_j =$ tr$(\mathbf{S})$ and is equal the total variance of the original variables. Therefore, the jth PC is response for a proportion P_j of total variation:

$$P_j = \frac{\lambda_j}{tr(\mathbf{S})}, \tag{3.3}$$

where the operator tr$(\cdot)$ denotes *trace* of matrix. So that the m first principal components, $m < q$, have a cumulative proportion $P^{(m)}$ of the total variation in the original data, where:

$$P^{(m)} = \frac{\sum_{j=1}^{m} \lambda_j}{\text{tr}(\mathbf{S})}$$

3.2 COVARIANCE AND CORRELATION MATRIX

Let $\mathbf{Z}$, the standardized matrix of the initial data matrix $\mathbf{X}$, in which the j-th column corresponds to the vector $\mathbf{z_j}$ with the n standardized observations of $\mathbf{x}_j$, then the covariance matrix of the set of $\mathbf{X}$ standardized is the $\mathbf{R}$ correlation matrix. Thus, we have the PCA correlation matrix method, and the PC $\mathbf{y}_k = \mathbf{Z}\boldsymbol{\beta}_k$ (EVERITT; HOTHORN, 2011).

If the PCs are extracted of matrix covariance, then the covariance and the correlation between variable jth and the kth is defined as, respectively:

$$Cov(x_k, y_j) = \lambda_j \beta_{jk}$$

$$r_{x_k,y_j} = \frac{\lambda_j \beta_{jk}}{\sqrt{Var(x_k)Var(y_j)}} = \frac{\lambda_j \beta_{jk}}{s_k\sqrt{\lambda_j}} = \frac{\sqrt{\lambda_j}\beta_{jk}}{s_k}$$

If the PCs are extracted of matrix correlation, the correlation coefficient between the variable j-th and the k-th PC is given by: $r = \sqrt{\lambda_k} ajk$

The percentage of variance is different for each PC, which is why pcs generated from the correlation matrix are required to represent the percentage of total variance. The trace of an $\mathbf{R}$ correlation matrix is equal to the number of p variables and therefore the total variance ratio of any PC is the variance of it divided by p. Therefore, when the data is at different scales or has very different variances, the correlation matrix is better suited to generate the PCs than the covariance matrix.

3.3 MINIMUM NUMBER OF PRINCIPAL COMPONENTS

Another important step to consider in principal component analysis is the minimum number of components. This amount should be obtained in a way that takes as much information as possible from the data, so there are several criteria in the literature that can be used (see Everitt and Hothorn (2011)). Here we use only one criterion: choose the components that present the percentage of the total cumulative variance between 70% and 90%. However, as the sample size increases, smaller values might be appropriate. In addition to this criterion, they describes other criteria that can be used to choose the components: those with eigenvalues greater than 0.7 (proposed by Jolliffe (1972)); from the analysis of the graph of λ_i *versus* i, introduced by Cattell (1966) and called scree diagram, in which the points are connected forming lines, and the last selected component is the one in which, from it, the line starts to have a little incline; etc.

4 INFLUENTIAL ANALYSIS

When the study involves dependent observations, the influence function includes the joint distribution of the data and it is not ideal that it derives from methods that remove or add observations. Genton and Ruiz-Gazen (2010) introduced the hair-plot, a tool to detect and visualize these observations in the context of dependent data based on the study of the effect of data perturbation on the estimators. For the development of this tool, they defined additive perturbations in the context of dependent data, allowing to obtain information about the behavior of these estimators considering different perturbation values. As an illustration, they use the pollution data by disturbing each observation and obtaining the method-of-moments sample variogram $2\hat{\gamma}(h)$. Genton and Ronchetti (2003) analyzed the same data and observed possible outliers in the residual values and observed that the highest residual referred to the value 40 in the location $(2, 2)$. Based on this, Genton and Ruiz-Gazen (2010) investigated the influence of each observation for the lags of distance $h = 1, 2, 3, 4$ and considering ζ a perturbation such that $\zeta \in [-40, 40]$. They noted that influential observation changes as distance increases.

4.1 EMPIRICAL, LOCAL AND ASYMPTOTIC INFLUENCE

Genton and Ruiz-Gazen (2010) proposed an additive perturbation structure so that it provides more information about how the estimator behaves by adding low perturbation values. Let $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)^\top$ be a vector data, and an additive perturbation of $\mathbf{Z}$, given by:

$$\mathbf{Z}[i, \zeta] = \mathbf{Z} + \zeta e_i, \tag{4.1}$$

where $e_i = 1$ for observation i and $e_i = 0$ otherwise, and $\zeta \in \mathbb{R}$ is a quantity of perturbation. For $\zeta = 0$, the value $\theta(\mathbf{Z})$ is preserved for all observations.

Let $\theta(\cdot)$ be the estimator of $\widehat{\theta}$ and $\theta(\mathbf{Z}[i, \zeta])$ the estimator of data perturbation in function $i = 1, \ldots, n$ and ζ. Genton and Ruiz-Gazen (2010) also showed that by plotting each of these estimators is possible to visualise the effect of influential observations. So the “hair-plot is a version of the empirical influence function with replacement". Still, they proposed two influential measures: local and asymptotic influential function. The local influence of the i-th observation is defined as:

$$\tau_i(\widehat{\theta}, \mathbf{Z}) = \frac{\partial}{\partial \zeta} \widehat{\theta}(\mathbf{Z}[i, \zeta])\bigg|_{\zeta=0}, \tag{4.2}$$

in such a manner that the largest absolute value of $\tau_i(\cdot)$ depicts the most influential observation.

On the other hand, the asymptotic influence of the i-th observation implies the influence of $\widehat{\theta}(\mathbf{Z})$ estimate when there is a high perturbation value for the i-th observation, and it is given by:

$$\nu_i(\widehat{\theta}, \mathbf{Z}) = \lim_{\zeta \to \infty} \widehat{\theta}(\mathbf{Z}[i, \zeta]), \tag{4.3}$$

4.2 INFLUENCE ON QUADRATIC FORM

In order to understand the effect of influence on data and to understand the structure and properties of the influence effect, Genton and Ruiz-Gazen (2010) defined the quadratic form of empirical, local and asymptotic influence as follows. Under a $Z[i, \zeta]$ contamination and considering the Equation 2.3, an influence quatratic form of semivariogram is given by:

$$\widehat{\gamma}^{(i,\zeta)}(h)_Q = \mathbf{Z}^\top \mathbf{A}(h)\mathbf{Z} + (\mathbf{Z}^\top \mathbf{A}(h)\mathbf{e}_i + \mathbf{e}_i^\top \mathbf{A}(h)\mathbf{Z})\zeta + (\mathbf{e}_i^\top \mathbf{A}(h)\mathbf{e}_i)\zeta^2,$$

and from that they obtained the expressions on quadratic form for the local and asymptotic influences from the functions 4.2 and 4.3, respectively, such as:

$$\tau_i(\widehat{\gamma}(h)_Q, \mathbf{Z}) = \mathbf{Z}^\top \mathbf{A}(h)\mathbf{e}_i + \mathbf{e}_i^\top \mathbf{A}(h)\mathbf{Z}$$

$$\nu_i(\widehat{\gamma}(h)_Q, \mathbf{Z}) = \infty.$$

In the same way, an influence quatratic form of cross-semivariogram (Equation 2.6) is given by:

$$\widehat{\gamma}_{\mathsf{uv}}^{(i,\zeta)}(h)_Q = \mathbf{Z}_{\mathsf{u}}^\top \mathbf{A}(h)\mathbf{Z}_{\mathsf{v}} + (\mathbf{Z}_{\mathsf{u}}^\top \mathbf{A}(h)\mathbf{e}_i + \mathbf{e}_i^\top \mathbf{A}(h)\mathbf{Z}_{\mathsf{v}})\zeta + (\mathbf{e}_i^\top \mathbf{A}(h)\mathbf{e}_i)\zeta^2,$$

and the quadratic form for the local and asymptotic influences, respectively, such as:

$$\tau_i(\widehat{\gamma}_{\mathsf{uv}}(h)_Q, \mathbf{Z}_{\mathsf{u}}, \mathbf{Z}_{\mathsf{v}}) = \mathbf{Z}_{\mathsf{u}}^\top \mathbf{A}(h)\mathbf{e}_i + \mathbf{e}_i^\top \mathbf{A}(h)\mathbf{Z}_{\mathsf{v}}$$

$$\nu_i(\widehat{\gamma}_{\mathsf{uv}}(h)_Q, \mathbf{Z}_{\mathsf{u}}, \mathbf{Z}_{\mathsf{v}}) = \infty.$$

4.3 GRAPHICAL TOOLS TO VISUALIZE INFLUENTIAL SPATIAL DATA

In this section, we describe the methodology for obtaining three of graphical tools developed in this master's thesis and we introduce them in order to identify influential observations in spatial data considering the semivariogram, cross-semivariogram and trace-semivariogram estimators denoted by $\widehat{\gamma}(\mathbf{h})$ (defined in section 2). First we generate a graph for the lag of distance h pair by pair using bihair-plot, then a graph is created for all lags of distance h considered using hair-plot introduced by (GENTON; RUIZ-GAZEN, 2010), but applying the principal component analysis (PCA) and finally we generated a hair-plot for functional data.

4.3.1 Bihair-plot

In order to analyze the influence two pairs of lags, we developed a graph called *bihair-plot*. Initially, a vector data $Z[i, \zeta]$ is obtained for i-th perturbed observation, and then a $\hat{\gamma}(h)$ value for each value of h. In the bihair-plot, a point $(\widehat{\gamma}(h_k), \widehat{\gamma}(h_l)), l \neq k, l, k = 1, \ldots, p$, is plotted for each ζ considered in the analysis, and it is connected so that it belong to the same observation and follow the order of the ζ's, forming a *curve*. Observation is considered influential if the curve related to it is closer to the diagonal and varies more in relation to the others.

4.3.2 Principal components hair-plot

In order to analyze the influence for all lags, the hair-plot was adapted to a version using principal component analysis (PCA). When we are in a multivariate problem, the PCA allows the interpretation and representation of the data in a graph, reducing its dimensionality through the principal component (PC) containing the maximum of data variability, which are linear combinations of the correlated variables of the data set. The PCs are not correlated with each other, and are ordered in such a way that the former contains most of the data variation. They are found through the decomposition of the centralized data matrix, reducing to a linear optimization problem, in such a way that minimizes the size of the data, maximizing the variability subject to certain restrictions. As the use of PCA is descriptive in nature, the observed data can be used regardless of their distribution. The assumption of normality can be assumed for inferential purposes (JOLLIFFE; CADIMA, 2016; EVERITT; HOTHORN, 2011).

Let a sampled data $Z = Z(\mathbf{s}) = (Z_1, Z_2, \ldots, Z_n)^\top$, where $\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$, and a perturbation of $\mathbf{Z}$ defined as: $\mathbf{Z}[i,\zeta]$. As for the perturbed data, we have a vector $\widehat{\gamma}(h)$ with n values related to $\mathbf{Z}[i,\zeta]$ for a given h. In this case, the following operations are performed:

$$Z[1,\zeta] = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} + \zeta \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} Z_1 + \zeta \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} \longrightarrow \widehat{\gamma}^{(1,\zeta)}(h)$$

$$\vdots$$

$$Z[n,\zeta] = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} + \zeta \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n + \zeta \end{pmatrix} \longrightarrow \widehat{\gamma}^{(n,\zeta)}(h)$$

Therefore, considering the vector $\mathbf{h}$ of size p, we obtain the following $n \times p$ matrix:

$$\check{\Gamma}(\mathbf{h}) = \begin{bmatrix} \widehat{\gamma}^{(1,\zeta)}(h_1) & \cdots & \widehat{\gamma}^{(1,\zeta)}(h_p) \\ \vdots & \ddots & \vdots \\ \widehat{\gamma}^{(n,\zeta)}(h_1) & \cdots & \widehat{\gamma}^{(n,\zeta)}(h_p) \end{bmatrix} \tag{4.4}$$

In this way, we want explain the maximum variation of $\widehat{\boldsymbol{\gamma}}(h_1), \ldots, \widehat{\boldsymbol{\gamma}}(h_p)$ through a new set of uncorrelated estimates, the PCs, $\widehat{\mathbf{y}}_1, \ldots, \widehat{\mathbf{y}}_p$, such that each $\widehat{\mathbf{y}}_k$, $k = 1, \ldots, p$, is a linear combination of q estimates $(q \leq p)$.

Commonly, the components are found in terms of the centralized matrix (JOLLIFFE; CADIMA, 2016). Thus, the first component $\widehat{\mathbf{y}}_1$ (PC1) has the following linear combination:

$$\widehat{\mathbf{y}}_1 = \beta_{11}\widehat{\boldsymbol{\gamma}}^\star(h_1) + \beta_{12}\widehat{\boldsymbol{\gamma}}^\star(h_2) + \ldots + \beta_{1p}\widehat{\boldsymbol{\gamma}}^\star(h_p) = \check{\Gamma}^\star(\mathbf{h})\boldsymbol{\beta}_1,$$

where $\boldsymbol{\beta}_1^\top\boldsymbol{\beta}_1 = 1$, Var($\widehat{\mathbf{y}}_1$) $= \boldsymbol{\beta}_1^\top\mathbf{S}\boldsymbol{\beta}_1$ is the sample variance of $\widehat{\mathbf{y}}_1$, $\mathbf{S}$ is the $n \times n$ sample covariance matrix of $\check{\Gamma}^\star(\mathbf{h})$, given by centred estimates $\boldsymbol{\gamma}^\star(h_k) = \gamma^{(i,\zeta)}(h_k) - \bar{\boldsymbol{\gamma}}(h_k), k = 1 \ldots, p$, as $\mathbf{S} = N^{-1}\check{\Gamma}^\star(\mathbf{h})^\top\check{\Gamma}^\star(\mathbf{h})$, and $\bar{\boldsymbol{\gamma}}(h_k)$ is the estimates mean of *k-th* column of $\check{\Gamma}^\star(\mathbf{h})$.

Similarly, the expression for the second component $\widehat{\mathbf{y}}_2$ (PC2) is given by

$$\widehat{\mathbf{y}}_2 = \beta_{21}\widehat{\boldsymbol{\gamma}}^\star(h_1) + \beta_{22}\widehat{\boldsymbol{\gamma}}^\star(h_2) + \ldots + \beta_{2p}\widehat{\boldsymbol{\gamma}}^\star(h_p) = \check{\Gamma}^\star(\mathbf{h})\boldsymbol{\beta}_2,$$

where $\boldsymbol{\beta}_2^\top\boldsymbol{\beta}_2 = 1$, and $\boldsymbol{\beta}_2^\top\boldsymbol{\beta}_1 = 0$, and Var($\widehat{\mathbf{y}}_2$) $= \boldsymbol{\beta}_2^\top\mathbf{S}\boldsymbol{\beta}_2$ is the sample variance of $\widehat{\mathbf{y}}_2$.

Hence, the jth PC is defined as $\widehat{\mathbf{y}}_j^\top = \check{\Gamma}^\star(\mathbf{h})\boldsymbol{\beta}_j$, where Var($\widehat{\mathbf{y}}_j$) $= \boldsymbol{\beta}_j^\top\mathbf{S}\boldsymbol{\beta}_j$ and requiring: $\boldsymbol{\beta}_j^\top\boldsymbol{\beta}_j = 1$ and $\boldsymbol{\beta}_j^\top\boldsymbol{\beta}_l = 0$ for $j < l, l = 1, \ldots, p$.

Then there is a problem of maximizing a function of multiple variables, subject to at least one constraint, therefore the method of *Lagrange multipliers* is used in such a way that β is the eigenvector, and S corresponding to this matrix's largest eigenvalue (see **Section** 3.1). Then, we want to maximizing $\boldsymbol{\beta}_j^\top\mathbf{S}\boldsymbol{\beta}_j - \lambda(\boldsymbol{\beta}_j^\top\boldsymbol{\beta}_j - 1)$, such that λ is a *Langrange multiplier*, finding λ from the equation:

$$\mathbf{S}\boldsymbol{\beta}_j = \lambda_j\boldsymbol{\beta}_j \Longleftrightarrow \mathbf{S}\boldsymbol{\beta}_j - \lambda_j\boldsymbol{\beta}_j = 0, \tag{4.5}$$

thus, the covariance between $\check{\Gamma}^\star(\mathbf{h})\boldsymbol{\beta}_j$ and $\check{\Gamma}^\star(\mathbf{h})\boldsymbol{\beta}_l$ is

$$\boldsymbol{\beta}_j^\top\mathbf{S}\boldsymbol{\beta}_l = \lambda_j\boldsymbol{\beta}_j^\top\boldsymbol{\beta}_l = \begin{cases} \lambda_j & j = l \\ 0 & (j \neq l). \end{cases} \tag{4.6}$$

The eigenvectors $\boldsymbol{\beta}_j$ and the linear combinations $\check{\Gamma}(\mathbf{h})\boldsymbol{\beta}_j$ are called *PC loadings* and *PC scores*, respectively.

From the expression 4.6, the total variance of the q principal components is given by: $\sum_{j=1}^{p}\lambda_j = \text{tr}(\mathbf{S})$ and is equal the total variance of the original estimates. Therefore, the j-th PC is response for a proportion P_j of total variation:

$$P_j = \frac{\lambda_j}{\text{tr}(\mathbf{S})}, \tag{4.7}$$

where the operator tr($\cdot$) means *trace*. So that the m first principal components, $m < q$, account for a proportion $P^{(m)}$ of the total variation in the original estimate, we have:

$$P^{(m)} = \frac{\sum_{j=1}^{m}\lambda_j}{\text{tr}(\mathbf{S})} \tag{4.8}$$

Another way to to choose β_j so as to find components with variations represented in the $\check{\Gamma}(\mathbf{h})$ estimates by following the steps below, defining the standardized weights that maximize Var$(\widehat{\mathbf{y}}_j)$:

1. find the weight vector $\boldsymbol{\xi}_1 = (\xi_{11}, \ldots, \xi_{1p})^\top$ for which values of

$$\widehat{y}_1^{(i)} = \sum_j \xi_{j1} \widehat{\gamma}^{(i)}(h_j) = \boldsymbol{\xi}_1^\top \widehat{\gamma}^{(i)}(\mathbf{h})$$

   have the highest possible quadratic mean: $N^{-1} \sum \widehat{y}_1^{(i)2}$ subject to $\xi_1^2 ds = 1$;

2. in the *m-th* step, find ξ_m with new values $\widehat{y}_m^{(i)} = \boldsymbol{\xi}_m^\top \widehat{\gamma}^{(i)}(\mathbf{h})$. Thus, ξ_m has a maximum quadratic average subject to $||\xi_m||^2 ds = 1$ and the $m-1$ restrictions: $\xi_k^\top \xi_m = 0, k < m$.

It is ideal to use the data correlation matrix $\mathbf{R}$ instead of the covariance $\mathbf{S}$, when $\widehat{\gamma}(\mathbf{h_j})$ are on very different scales or different variances (EVERITT; HOTHORN, 2011). In this case, the eigenvalues λ are obtained by finding the root of the following equation:

$$\text{det}(\mathbf{R} - \lambda \mathbf{I}) = 0.$$

Then, the eigenvectors can found solving Equation 4.5, using $\mathbf{R}$ instead $\mathbf{S}$.

The next step is to find the minimum number of principal components that can be used in the analysis. As mentioned earlier (see Chapter 3), we choose the components that present at least 70% of the cumulative variance of the estimates.

5 APPLICATIONS

In order to illustrate the methods and visual tools proposed and presented in the previous sections, we apply them to pollution data in section 5.1, to meuse data from the sp package in R in section 5.2 and maritimes data from the geofd package in R in section 5.3.

5.1 POLLUTION DATA

The sample of 100 pollution data, comes from a 9×9 regular grid, consists of reflectance values from the pumping of waste material into the English Channel. Genton and Ruiz-Gazen (2010) analyzed it in order to model the possible dependence structure of the pollution levels, so that high pollution levels induce high reflectance values, with the previous removal of the linear trend producing residuals $Z(s_1, s_2), s_1, s_2 = 1, \ldots, 9$. Then they computed the empirical method-of-moments variogram $2\hat{\gamma}(h)$ estimator assuming isotropy. Considering the data modified by the additive perturbation $\zeta \in [-40 : 40]$, they investigated the empirical influence of each residual on the sample variogram for the lags $h = \{1, 2, 3, 4\}$ through the hair-plot and found observation $\#17$, located in $(2, 2)$, as an influential point that corresponds to the maximum reflectance residual value of the data equal to 40. In addition, in the present work we studied the empirical influence pairing the sample semivariogram $\widehat{\gamma}^{(i,\zeta)}(h)$, obtained by the perturbed data $Z[\zeta, i], i, \ldots, 100$ and illustrated on the bihair-plot. Descriptive statistics for the estimates are shown in the Table 2, in which we observed that the data range from -26.70 to 40, and the observations associated with their maximum values are presented for $h = \{1, 2, 3, 4\}$. It was possible to observe that the influential observations change, in such a way that as the lag grows, observation $\#59$ becomes influential, initially being observation $\#17$ considering small distances.

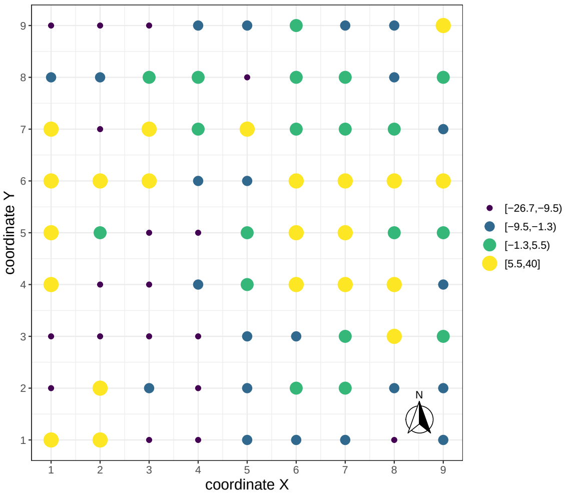
Tabela 1 – Descriptive statistics values of pollution data.

Statistics	**reflectance**
Minimum	-26.70
1st Quartile	-9.50
Median	-1.30
Mean	-1.39
3rd Quartile	5.50
Maximum	40.00

Source: Elaborated by the author (2021).

Figure 7 describes the distribution of the data through quartile intervals. Values above the third quartile (last interval) correspond to the maximum values of the sample, with 40 being the highest value, at location $(2, 2)$.

Figura 1 – Distribution of reflectance values by quartile interval - pollution data.



Source: Elaborated by the author (2021).

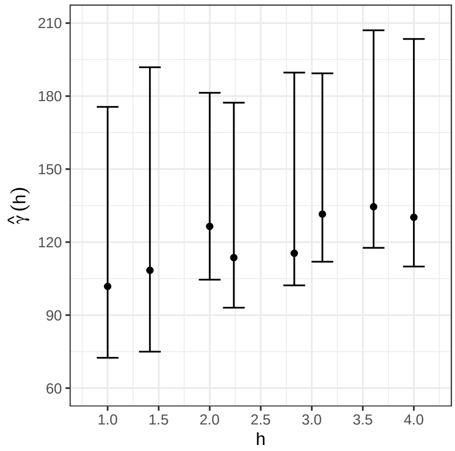
The $\widehat{\gamma}^{(i,\zeta)}(h)$ amplitude also was shown in Figure 2 for each h, wherein the original sample semivariogram, when $\zeta = 0$, was represented by points.

Tabela 2 – Descritive statistics of method-of-moments semivariogram and the observation corresponding to the maximum value of the estimates for each lag ($\zeta > 0$).

Statistics	$\hat{\gamma}(1)$	$\hat{\gamma}(2)$	$\hat{\gamma}(3)$	$\hat{\gamma}(4)$	$\hat{\gamma}(5)$	$\hat{\gamma}(6)$	$\hat{\gamma}(7)$	$\hat{\gamma}(8)$
Minimum	72.46	74.97	104.6	93.04	102.2	111.9	117.6	110.0
Median	107.03	112.91	131.1	118.16	119.5	135.8	138.5	134.6
Mean	109.69	116.31	134.4	121.51	123.3	139.4	142.4	138.1
Maximum	175.57	191.85	181.4	177.29	189.7	189.4	207.0	203.5
Observation	17	17	59	59	25	59	59	59

Source: Elaborated by the author (2021).

Figura 2 – Semivariogram amplitude graph for perturbed data represented by the bars, where the minimum and the maximum is the less and highest value of $\widehat{\gamma}(h)$ for $\zeta = [-40, 40]$, respectively, and semivariogram for original data represented by the points where $\zeta = 0$ - pollution data



Source: Elaborated by the author (2021).

The Figure 3 shows the largest residual, when $\zeta > 0$, corresponding to observation $\#17$ and it is identified by the red curve and black curve for $h = \{1, 2, (1, 2), (1, 3), (1, 4), (2, 2), (2, 3), (3, 4)\}$ and $h = \{3, 4, (3, 4)\}$, respectively, where $(\cdot, \cdot)$ is the pair of lags h related to estimates represented on bihair-plot. For $h = \{3, 4, (3, 4)\}$, observation $\#59$ (location (7,5)) is the most influential ($\zeta > 0$), since its curve corresponds to the most extreme.

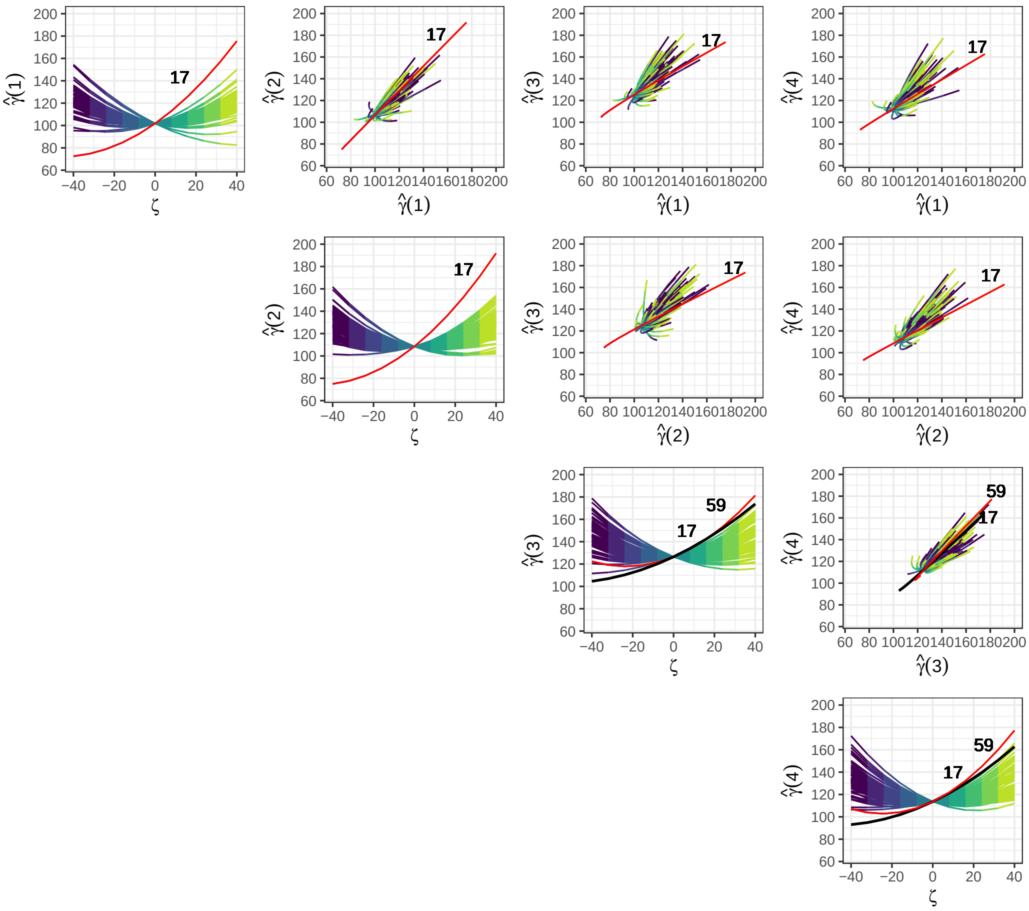
When the data is perturbed, and as the data are spatially dependent, it will have different scales for each lag h, as seen in the Figure 2. So the ideal is to get the PCs using the sample correlation matrix of $\check{\mathbf{\Gamma}}^{\star}(\mathbf{h})$. Therefore, we obtained the PC *scores* from the correlation matrix of the sample semivariogram, as defined in equation 4.5 (section 4.3.2), first for $\mathbf{h} = \{1, 2, 3, 4\}$ and then taking into account all $\mathbf{h}$. The first component $\widehat{\mathbf{y}}_1$ presented PC *loadings* equal to 0.5 for all lags $\mathbf{h}$, giving equal importance to the values of the semivariogram independent of the distance. The PC *loadings* were different for the second component $\widehat{\mathbf{y}}_2$, so that shortest distance ($h = 1$) showed the highest PC *loading*.

$$\widehat{\mathbf{y}}_1 = 0.50\widehat{\boldsymbol{\gamma}}^{\star}(1) + 0.50\widehat{\boldsymbol{\gamma}}^{\star}(2) + 0.50\widehat{\boldsymbol{\gamma}}^{\star}(3) + 0.50\widehat{\boldsymbol{\gamma}}^{\star}(4)$$

$$\widehat{\mathbf{y}}_2 = 0.58\widehat{\boldsymbol{\gamma}}^{\star}(1) + 0.42\widehat{\boldsymbol{\gamma}}^{\star}(2) + -0.51\widehat{\boldsymbol{\gamma}}^{\star}(3) + -0.48\widehat{\boldsymbol{\gamma}}^{\star}(4)$$

where: $\boldsymbol{\gamma}^{\star}(h_k) = \gamma^{(i,\zeta)}(h_k) - \bar{\boldsymbol{\gamma}}(h_k), k = 1, 2, 3, 4$, where $\bar{\boldsymbol{\gamma}}(h_k)$ is the estimates mean of *k-th* column of $\check{\Gamma}^{\star}(\mathbf{h})$.

Figura 3 – Hair-plots in the principal diagonal and the bihair-plot on top of the sample semivariogram on the reflectance residual values for spatial lag distance $h = 1, 2, 3, 4.$ and $\zeta \in (-40, 40)$, for the pollution data



Source: Elaborated by the author (2021).

Considering all the lags $\mathbf{h}$, the first component also presented similar PC *loadings*, whereas in the second component, they had different values, where the greatest PC *loading* is for $h = 1$.

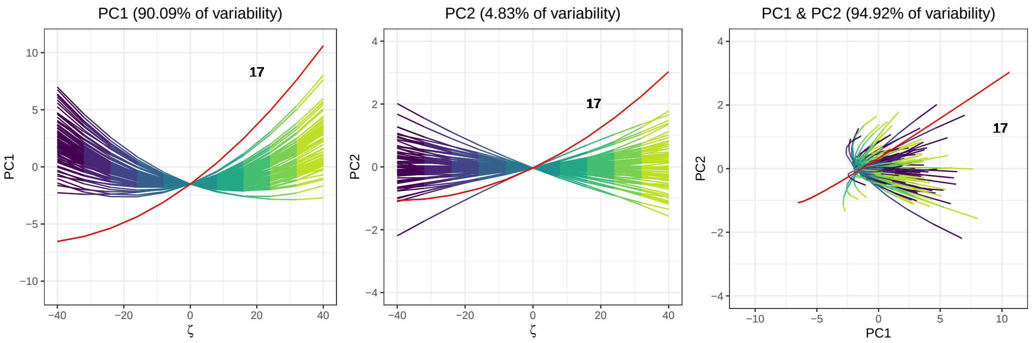
$$\widehat{\mathbf{y}}_1 = 0.34\widehat{\gamma}^{\star}(1)+0.35\widehat{\gamma}^{\star}(2)+0.36\widehat{\gamma}^{\star}(3)+0.37\widehat{\gamma}^{\star}(4)+0.35\widehat{\gamma}^{\star}(5)+0.37\widehat{\gamma}^{\star}(6)+0.35\widehat{\gamma}^{\star}(7)+0.34\widehat{\gamma}^{\star}(8)$$

$$\widehat{\mathbf{y}}_2 = 0.59\widehat{\gamma}^{\star}(1) + 0.46\widehat{\gamma}^{\star}(2) + 0.14\widehat{\gamma}^{\star}(3) - 0.46\widehat{\gamma}^{\star}(5) - 0.13\widehat{\gamma}^{\star}(6) - 0.43\widehat{\gamma}^{\star}(7)$$

When generating the hair-plot and the bihairplot of the principal components, the most influential observation was still $\#17$. Figure 4 illustrates the hair-plot for the first PC *score*, containing 90.0% of variability of the sample semivariogram, the second PC *score*, with 4.83% of variability, and the bihair-plot is generated for the two components and contains 94.92%

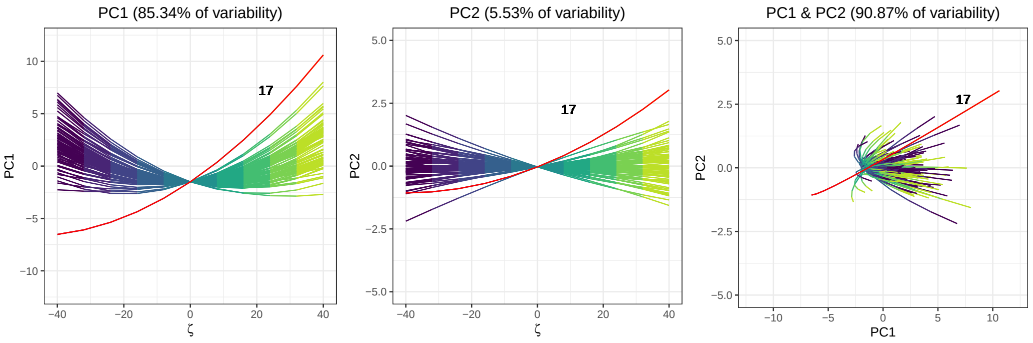
of sample semivariogram cumulative variability, for $h = 1, \ldots, 4$. The hair-plots and the bihair-plot considering all the values of $\mathbf{h}$, shown in the Figure 5, have the same behavior, in which the first component presents 85.34% of variability and the second presents 5.53% of variability, so that the first two components contained 90.87% of cumulative variability. For both scenarios, $h = 1, \ldots, 4$ and all $\mathbf{h}$, the first component presented more than 70% of the sample semivariogram variability, so that only the first component can be used to represent the estimate of all lags, and is most ideal since PC *loadings* have similar positive weights for all lags $\mathbf{h}$, thus giving importance to them regardless of distance.

Figura 4 – From left to right: PC hair-plots for PC1 and PC2 with 90.09% and 4.83% of sample semivariogram variability, respectively, and PC bihair-plot crossing PC1 and PC2 containing 94.92% of sample semivariogram cumulative variability, considering spatial lag distance $h = 1, 2, 3, 4.$ and $\zeta \in (-1, 1)$ - pollution data



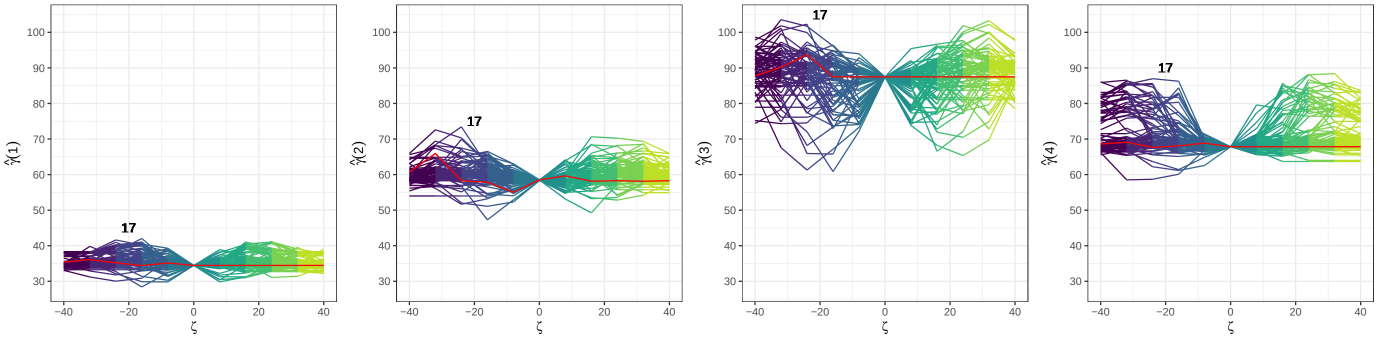
Source: Elaborated by the author (2021).

Figura 5 – From left to right: PC hair-plots for PC1 and PC2 with 87.88% and 5.75% of sample semivariogram variability, respectively, and PC bihair-plot crossing PC1 and PC2 containing 93.63% of sample semivariogram cumulative variability, considering spatial lag distance $h = 1, \ldots, 8$ and $\zeta \in (-1, 1)$ - pollution data



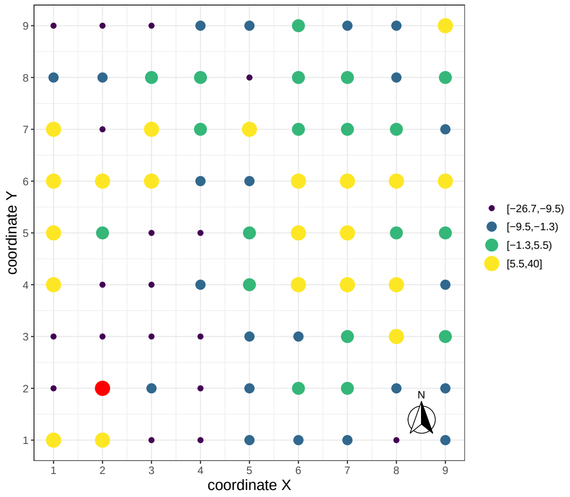
Source: Elaborated by the author (2021).

Figura 6 – Hair-plots of the MVE sample semivariogram for the reflectance residual for $h = 1, 2, 3, 4$ - pollution data



Source: Elaborated by the author (2021).

Figura 7 – Distribution of reflectance values by quartile interval. The red point on the grid indicates the most influential point, corresponding to the maximum data value (observation #17) - pollution data.

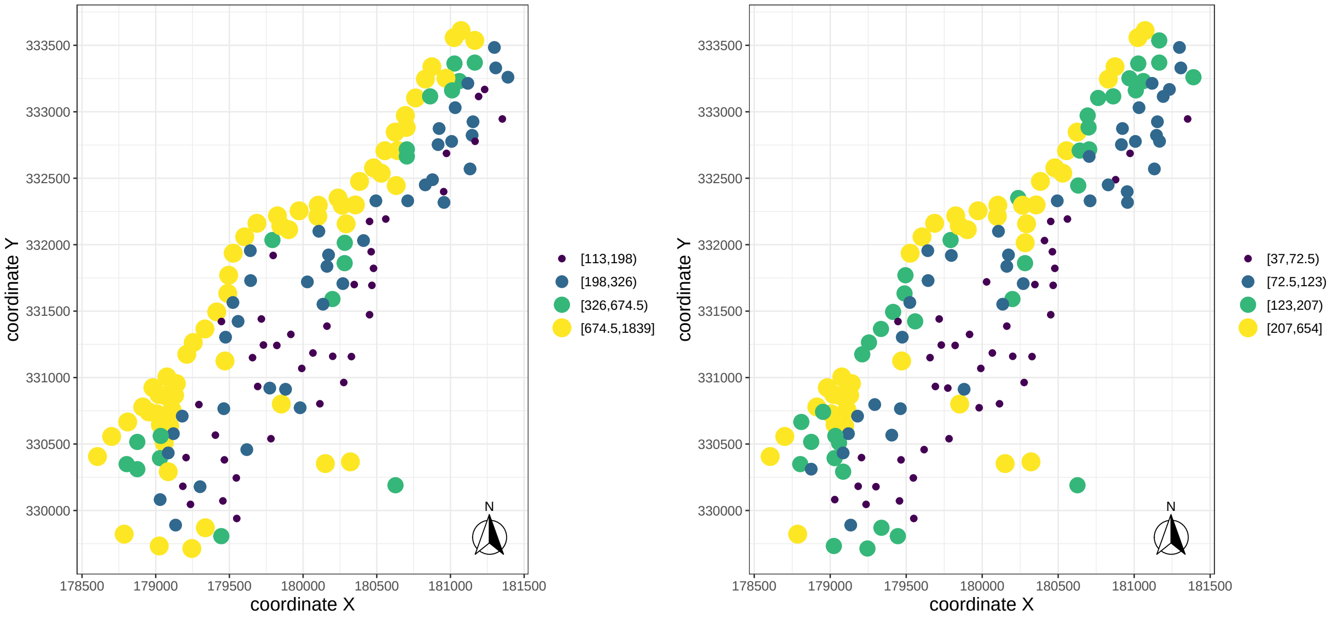


Source: Elaborated by the author (2021).

5.2 MEUSE DATA

The floodplains of the Meuse River contain large amounts of heavy metals due to the decomposition of contaminated sediments accumulated on the river bank. The meuse data set from the sp package is composed of variables referring to concentrations of different types of heavy metals (in pmm), with 155 observations, and their respective collection sites, which is the upper layer of the ground of the Meuse floodplain, near the village of Stein (NL). The sample was collected from an area of approximately $15m \times 15m$ (PEBESMA; BIVAND, 2005; BIVAND; PEBESMA; GOMEZ-RUBIO, 2013). Here, we studied two metals: zinc $(\mathbf{Z}_u)$ and lead $(\mathbf{Z}_v)$ concentrations. First, we applied $\log(\cdot)$ in order to normalize the data, and then we study the spatial variability of zinc and lead separately, from the sample semivariogram, and the spatial variability associating the two metals through the cross-semivariogram. In this case, we expect that where there is a higher concentration of zinc, there will also be a higher concentration of lead. In Figure 8, we can see that the highest concentration of metal is found on the banks of the river, in addition to being the highest values of metals (belonging to the third interval).

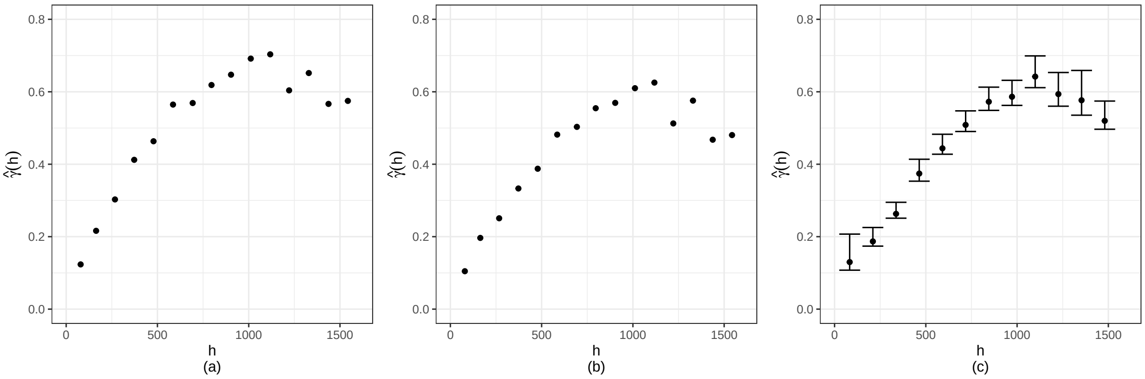
Figura 8 – Distribution map of zinc and lead values by quartile interval from left to right, respectively - meuse data.



Source: Elaborated by the author (2021).

In Figure 9(a) and (b), we can see that there is spatial dependence, because the semivariogram value increases when h becomes larger. In Figure 9(c) we identified a positive spatial association between zinc and lead values trough the original cross-semivariogram $\widehat{\gamma}(h)$ represented by points and we analysed $\check{\Gamma}(h)$ amplitude of each h, that is, each column $\widehat{\gamma}(h)$ of the $\check{\Gamma}(\mathbf{h})$ matrix, represented by bars. We observed that the scale of the estimates varies for each h.

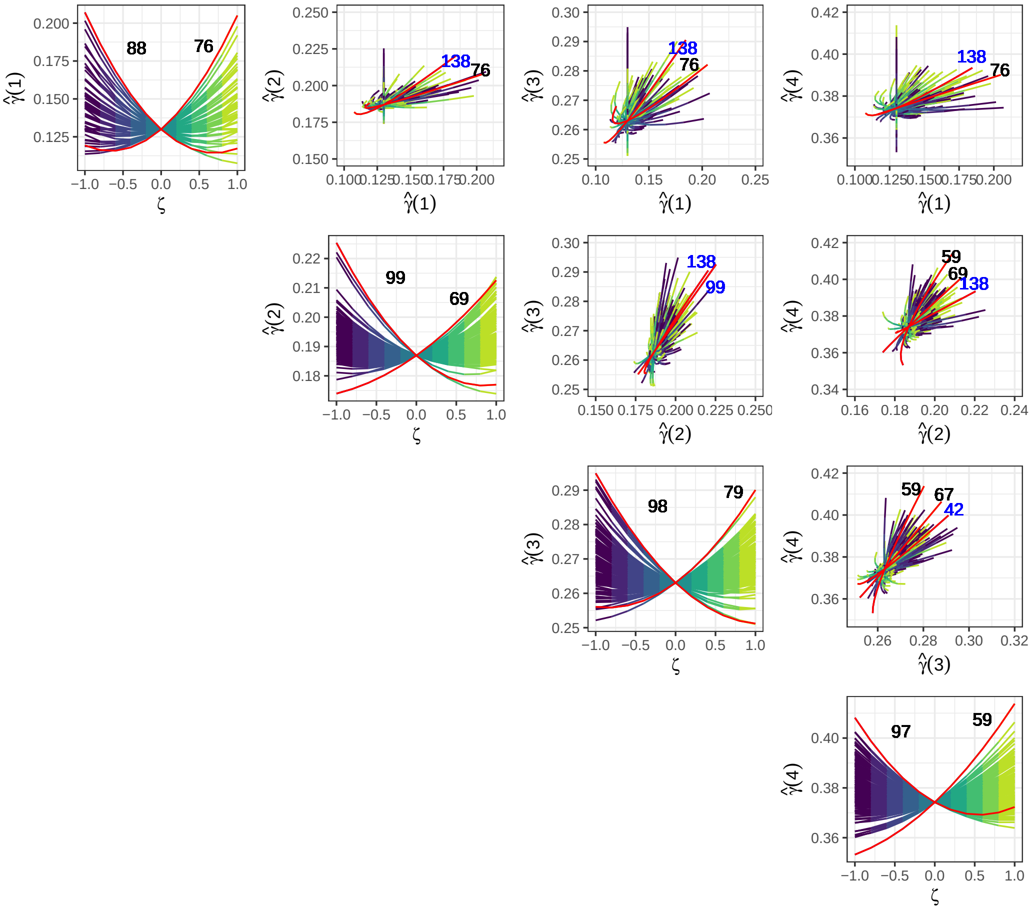
Figura 9 – Estimates of method-of-moments semivariogram for (a) *log* of zinc and (b) *log* of lead, (c) original sample cross-semivariogram for *log* of zinc and *log* of lead represented by the points and cross-semivariogram amplitude graph of for *log* of zinc and *log* of lead perturbed, where the minimum and the maximum is the less and highest value of $\widehat{\gamma}(h)$ for $\zeta = [-1, 1]$, respectively, represented by the bars - meuse data



Source: Elaborated by the author (2021).

Before obtaining the components, we investigated for $h = 1, 2, 3, 4$, just like it was done in section 5.1, but now we consider $\zeta \in [-1, 1]$. In Figure 10, the sample cross-semivariogram on original data refers to the value in the graph where $\zeta = 0$. The diagonal graphs correspond to the hair-plot generated for each lag h, and those on the top diagonal, being the same as those on the bottom diagonal, refer to the bihair-plot. The curve of the most influential point is highlighted in red. Observing the hair-plot for $h = 1$ and $\zeta > 0$, we have that observation $\#76$ stands out, with values 680 ppm and 241 ppm (zinc and lead, respectively) at location $179095, 330636$ in meters on Netherlands topographical, and, from the bihair-plot, we see that it becomes less influential as h increases. When analyzing for $h = \{3, 4, (3, 4)\}$, we have that $\#67$ ($\zeta > 0$) becomes more influential. For the pairs of lags $h = \{(2, 3); (2, 4); (3, 4)\}$ and $\zeta > 0$ we can see that $\#59$ is the most influential point. Descriptive statistics for the estimates are shown in the Table 3, and the observations associated with their maximum values are presented for $h = 1, 2, 3, 4$.

Figura 10 – Hair-plots in the principal diagonal and the bihair-plot on top diagonal of the sample cross-semivariogram on the values of zinc and lead, for spatial lag distance $h = 1, 2, 3, 4.$ and $\zeta \in (-1, 1)$. Blue values in the bihair-plot indicate the curve for which $\zeta < 0$. - meuse data



Source: Elaborated by the author (2021).

Tabela 3 – Descritive statistics of method-of-moments cross-semivariogram and the observation corresponding to the maximum value of the estimates for each lag $(\zeta > 0)$ - meuse data.

Statistics	$\hat{\gamma}(1)$	$\hat{\gamma}(2)$	$\hat{\gamma}(3)$	$\hat{\gamma}(4)$
Minimum	0.11	0.17	0.25	0.35
Median	0.13	0.19	0.26	0.38
Mean	0.13	0.19	0.26	0.37
Maximum	0.21	0.23	0.29	0.41
Observation	76	74	79	59

Source: Elaborated by the author (2021).

In PC hair-plot, the idea is to investigate whether there are any influential points considering the spatial dependence. By applying PCA for the $\check{\Gamma}(h)$ and considering $h = 1, 2, 3, 4$, we have that the first two components $\widehat{\mathbf{y}}_1$ and $\widehat{\mathbf{y}}_2$ (PC1 and PC2, respectively) carry 85.85% of the variability of the estimates and we obtain the following expressions for them:

$$\widehat{\mathbf{y}}_1 = 0.34\widehat{\gamma}^\star(1) + 0.54\widehat{\gamma}^\star(2) + 0.57\widehat{\gamma}^\star(3) + 0.52\widehat{\gamma}^\star(4),$$

$$\widehat{\mathbf{y}}_2 = 0.90\widehat{\gamma}^\star(1) - 0.19\widehat{\gamma}^\star(3) - 0.39\widehat{\gamma}^\star(4).$$

The first component $\widehat{\mathbf{y}}_1$ presented similar PC *loadings*, except for $h = 1$, which presented the lowest value equal to 0.34 for all lags $\mathbf{h}$, giving similar importance to the values of the cross-semivariogram. The PC *loadings* were different for the second component $\widehat{\mathbf{y}}_2$, so that shortest distance ($h = 1$) showed the highest PC *loading* equal 0.9.

Considering all lags h, PC $loadings$ were positive also for $\widehat{\mathbf{y}}_1$, where the lower value 0.10 was referring to $h = 1$. The PC *loading* for $\widehat{\mathbf{y}}_2$ varied for all lags, such that it presented higher values for $h = \{1, 2\}$.

$$\widehat{\mathbf{y}}_1 = 0.10\widehat{\gamma}^\star(1) + 0.22\widehat{\gamma}^\star(2) + 0.27\widehat{\gamma}^\star(3) + 0.29\widehat{\gamma}^\star(4) + 0.30\widehat{\gamma}^\star(5) + 0.32\widehat{\gamma}^\star(6) + 0.32\widehat{\gamma}^\star(7) +$$

$$0.32\widehat{\gamma}^\star(8) + 0.32\widehat{\gamma}^\star(9) + 0.31\widehat{\gamma}^\star(10) + 0.30\widehat{\gamma}^\star(11) + 0.30\widehat{\gamma}^\star(12),$$

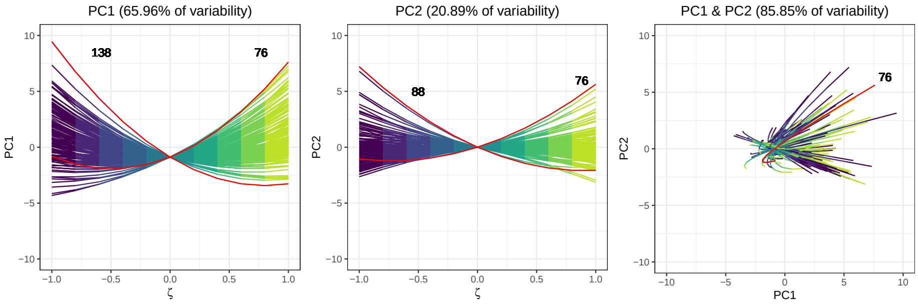
$$\widehat{\mathbf{y}}_2 = 0.59\widehat{\gamma^\star}(1) + 0.50\widehat{\gamma}^\star(2) + 0.39\widehat{\gamma}^\star(3) + 0.18\widehat{\gamma}^\star(4) - 0.11\widehat{\gamma}^\star(7) - 0.16\widehat{\gamma}^\star(8) - 0.20\widehat{\gamma}^\star(9) -$$

$$0.23\widehat{\gamma}^\star(10) - 0.22\widehat{\gamma}^\star(11) - 0.18\widehat{\gamma}^\star(11).$$

In Figure 12, we can see that PC1 has $67,83\%$ of the data variability, and the most influential points were $\#68$ ($\zeta < 0$) and $\#54$ ($\zeta > 0$). The observation $\#54$ correspond to the maximum values of the zinc and lead variables equal to 1839 and 654 at location (179973, 332255) in meters, respectively. PC2, on the other hand, has $11,21\%$ of the data variability, and the most influential points were $\#138$ and $\#76$ for $\zeta < 0$ and $\zeta > 0$, in due order. While PC1 loads information from the estimates of all lags, PC2 loads information from $\widehat{\gamma}(1), \widehat{\gamma}(3), \widehat{\gamma}(4)$, beeing $\widehat{\gamma}(1)$ with greater weight.

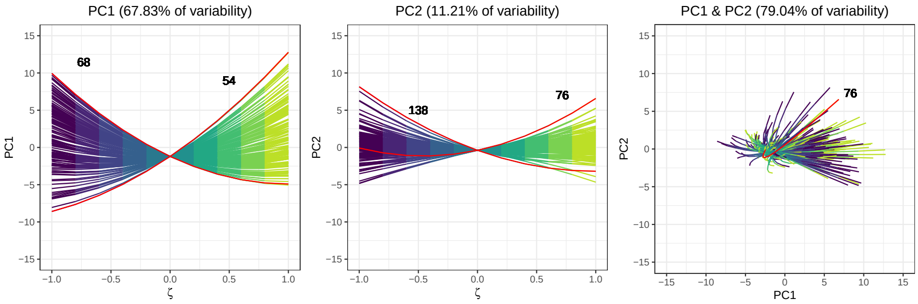
In Figure 6, it was observed that, from the cross semivariogram MVE, it is not possible to detect influencing points, that is, the estimator was not affected by the additive perturbations ζ such that the values of the estimates fluctuated and did not show an increasing behavior or decreasing as $\zeta \longrightarrow \infty$.

Figura 11 – From left to right: PC hair-plots for PC1 and PC2 with $65,96\%$ and $20,89\%$ of the sample cross-semivariogram variability, respectively, and PC bihair-plot crossing PC1 and PC2 containing 85.85% of the sample cross-semivariogram cumulative variability, considering lags distance $h = 1, 2, 3, 4.$ and $\zeta \in (-1, 1)$ - meuse data



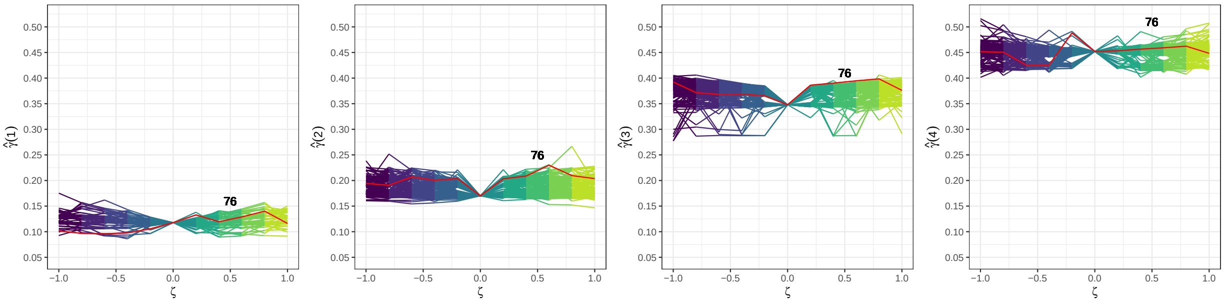
Source: Elaborated by the author (2021).

Figura 12 – From left to right: PC hair-plots for PC1 and PC2 with 67.83% and 11.21% of the sample cross-semivariogram variability, respectively, and PC bihair-plot crossing PC1 and PC2 containing 79.04% of the sample cross-semivariogram cumulative variability, considering lags distance $h = \{1, \ldots, 12\}$ and $\zeta \in (-1, 1)$ - meuse data



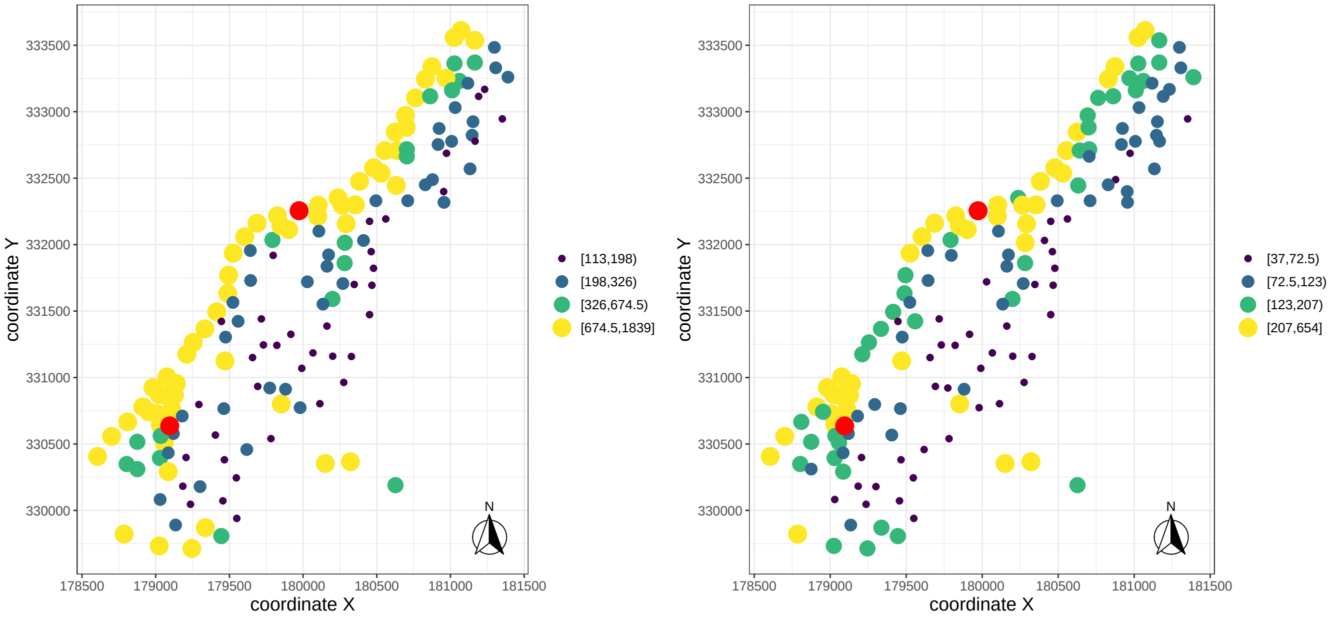
Source: Elaborated by the author (2021).

Figura 13 – Hair-plots of the MVE sample cross-semivariogram for the $\log$ of zinc and $\log$ of lead for $h = 1, 2, 3, 4$ - meuse data



Source: Elaborated by the author (2021).

Figura 14 – Distribution map of zinc and lead values by quartile interval from left to right, respectively. The points in red correspond to the most influential observation $(\#54)$ and a possible influential point $(\#76)$ $(\zeta > 0)$, at locations $(179973, 332255)$ and $(179095, 330636)$ given in meters. - meuse data.



Source: Elaborated by the author (2021).

5.3 MARITIMES DATA

Maritimes data are a large number of measurements corresponding to temperature records in 35 weather stations located in Canadian Maritime Provinces, named as Nova Scotia (NS), New Brunswisk (NB), and Prince Edward Island (PEI). Giraldo, Mateu and Delicado (2012) analyzed the average temperatures between the 1960s and 1994s for each station, which were initially obtained by the Meteorological Service of Canada. The data belongs to the geofd package (maritimes.data), and it also contains the average of values between the stations (maritimes.average) and their respective coordinates latitude and logitude (maritimes.coord). This information was crossed, and the trace-semivariogram hair estimates were obtained. Within the function, the data were smoothed from the nonparametric B-spline function, and trace-semivariogram estimates were generated from a vector of perturbations ranging from -1 to 1, presented in the Table 4, together with the observations associated with the maximum estimated value of $h = \{1, 2, 3, 4\}$, where hair-plots and bihair-plots are generated and illustrated in the Figure 19. It was observed that two observations are detected for $h = 3$ ($\#18$ and $\#19$) from the maximum value of the estimates, and that is why it is important to evaluate the hair-plot.

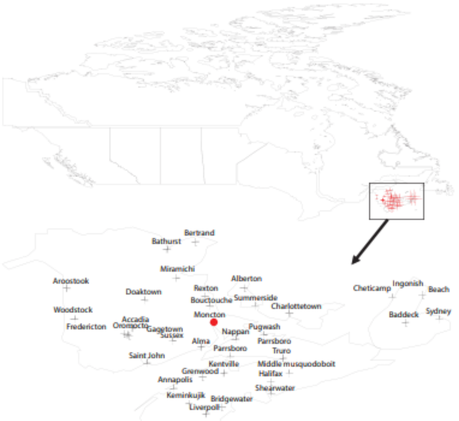
Tabela 4 – Descritive statistics of method-of-moments trace-semivariogram and the observation corresponding to the maximum value of the estimates for each lag ($\zeta > 0$) - pollution data.

Statistics	$\hat{\gamma}(1)$	$\hat{\gamma}(2)$	$\hat{\gamma}(3)$	$\hat{\gamma}(4)$
Minimum	217.9	370.9	644.2	830.6
Median	265.9	423.5	680.7	905.8
Mean	290.0	449.8	701.9	936.3
Maximum	570.1	745.7	924.3	1304.0
Observation	16	22	18 and 19	22

Source: Elaborated by the author (2021).

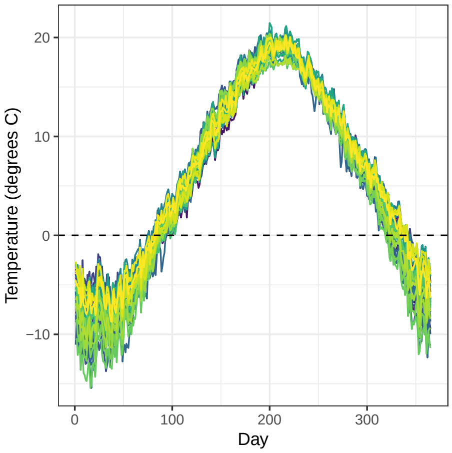
The data was modified by the additive perturbation $\zeta \in [-1 : 1]$, we want to identify which station presented temperature values such that the non-parametric function smoothed to disturbed data is influential. In this work, it was studied the empirical influence pairing the sample trace-semivariogram $\hat{\gamma}^{(i,\zeta)}(h)$, obtained by perturbed data $Z[\zeta, i], i, \ldots, 35$ and illustrated on the hairplot and bihair-plot. Figure 18 shows the $\hat{\gamma}^{(i,\zeta)}(h)$ amplitude also for each h, and the original sample trace-semivariogram, when $\zeta = 0$, was represented by points. Figure 17 illustrates the curves obtained using a base B-spline with 65 functions.

Figura 15 – Averages location of daily temperature curves observed at 35 weather stations of the Canadian Maritime provinces - maritimes data.



Source: Giraldo, Mateu and Delicado (2012).

Figura 16 – Matplot of daily temperature curves observed at 35 weather stations of the Canadian Maritime provinces - maritimes data



Source: Elaborated by the author (2021).

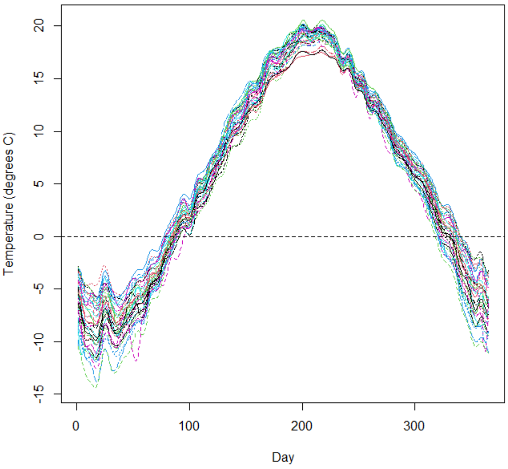
It was investigated whether there are any influential function on maritimes data. By applying PCA for the $\check{\Gamma}(h)$ and considering $h = 1, 2, 3, 4$, the first two components $\widehat{\mathbf{y}}_1$ and $\widehat{\mathbf{y}}_2$ (PC1 and PC2, respectively) carried 96.43% of the variability of the estimates and we obtain the following expressions for them:

$$\widehat{\mathbf{y}}_1 = 0.39\widehat{\gamma}^{\star}(1) + 0.54\widehat{\gamma}^{\star}(2) + 0.54\widehat{\gamma}^{\star}(3) + 0.51\widehat{\gamma}^{\star}(4),$$

$$\widehat{\mathbf{y}}_2 = 0.85\widehat{\gamma}^{\star}(1) - 0.21\widehat{\gamma}^{\star}(3) - 0.47\widehat{\gamma}^{\star}(4).$$

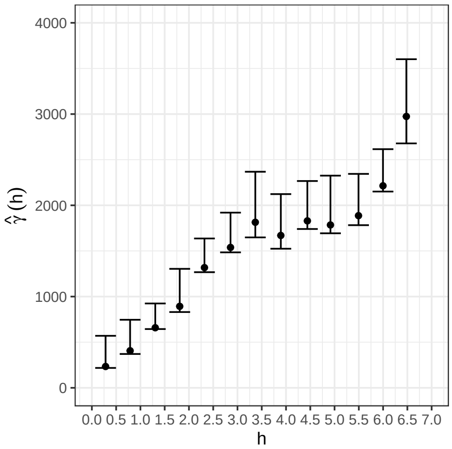
As seen for the trace-semivariogram estimates, the first component $\widehat{\mathbf{y}}_1$ presented similar PC *loadings*, except for $h = 1$ (the weight was equal to 0.39) for all lags $\mathbf{h}$, giving similar

Figura 17 – Smoothed data of daily temperature curves obtained by B-spline with 65 basis functions - maritimes data.



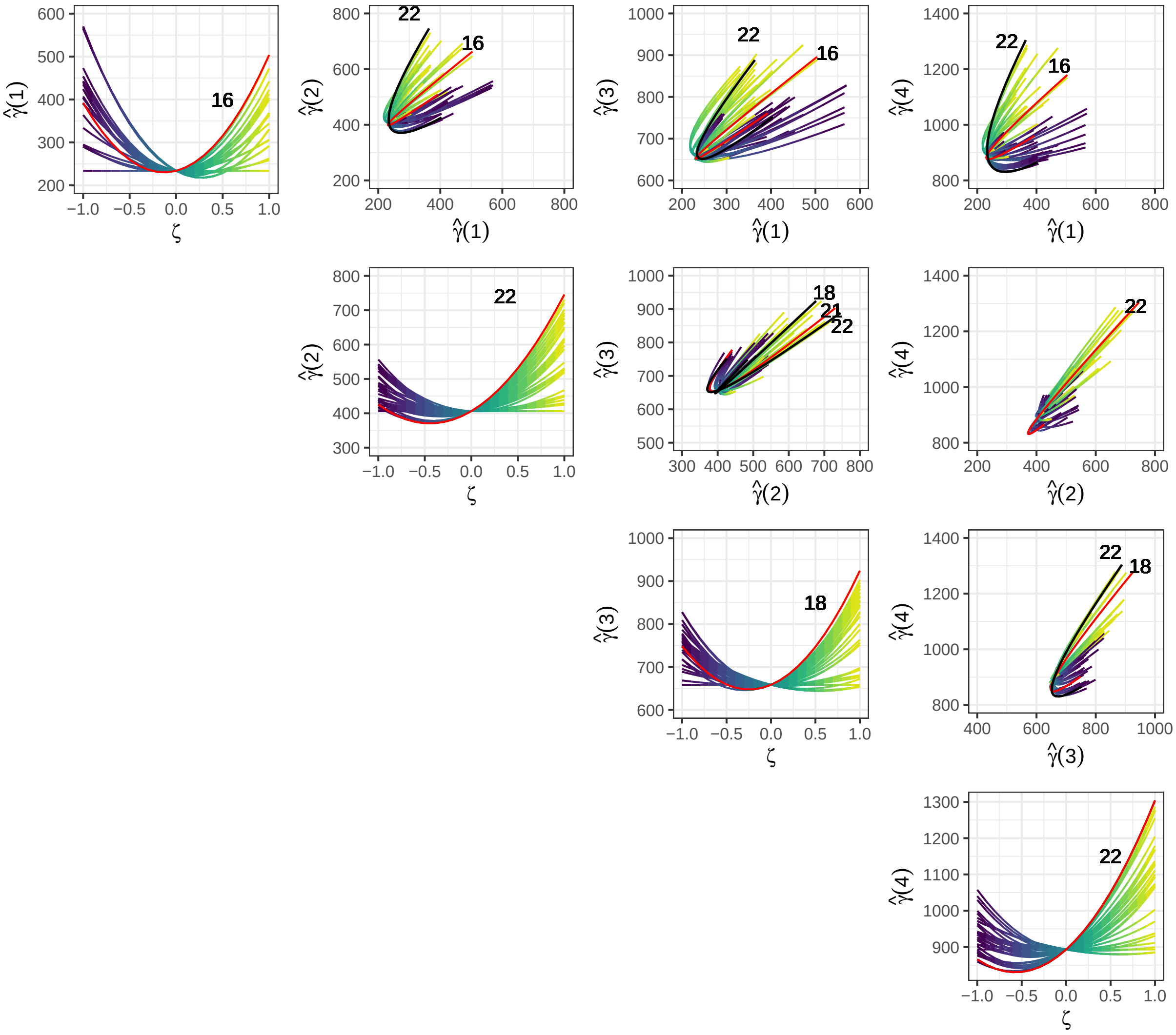
Source: Elaborated by the author (2021).

Figura 18 – Trace-semivariogram amplitude graph for perturbed data represented by the bars, where the minimum and the maximum is the less and highest value of $\widehat{\gamma}(h)$ for $\zeta = [-1, 1]$, respectively - maritimes data



Source: Elaborated by the author (2021).

Figura 19 – Hair-plots in the principal diagonal and the bihair-plot on top diagonal of the sample trace-semivariogram on the values of maritimes data, for spatial lag distance $h = 1, 2, 3, 4.$ and $\zeta \in (-1, 1)$. The black curve represents the observation that was influential by analyzing a lag among those taking part in the bihairplot - maritimes data



Source: Elaborated by the author (2021).

importance to the values of the trace-semivariogram. The PC *loadings* were different for the second component $\widehat{\mathbf{y}}_2$, so that shortest distance $(h = 1)$ showed the highest PC *loading* equal 0.85.

$$\widehat{\mathbf{y}}_1 = 0.27\widehat{\gamma}^{\star}(1) + 0.26\widehat{\gamma}^{\star}(2) + 0.30\widehat{\gamma}^{\star}(3) + 0.25\widehat{\gamma}^{\star}(4) + 0.34\widehat{\gamma}^{\star}(5) + 0.25\widehat{\gamma}^{\star}(6) + 0.31\widehat{\gamma}^{\star}(7) +$$

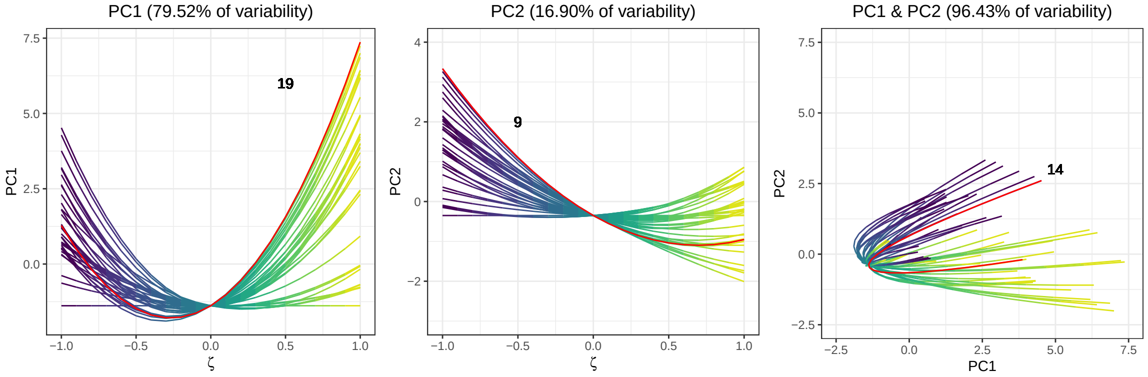
$$0.32\widehat{\gamma}^{\star}(8) + 0.25\widehat{\gamma}^{\star}(9) + 0.23\widehat{\gamma}^{\star}(10) + 0.28\widehat{\gamma}^{\star}(11) + 0.16\widehat{\gamma}^{\star}(12),$$

$$\widehat{\mathbf{y}}_2 = -0.35\widehat{\gamma}^{\star}(2) - 0.21\widehat{\gamma}^{\star}(3) - 0.39\widehat{\gamma}^{\star}(4) - 0.14\widehat{\gamma}^{\star}(5) + 0.36\widehat{\gamma}^{\star}(7) -$$

$$0.11\widehat{\gamma}^{\star}(8) + 0.16\widehat{\gamma}^{\star}(9) + 0.34\widehat{\gamma}^{\star}(10) + 0.38\widehat{\gamma}^{\star}(11) - 0.13\widehat{\gamma}^{\star}(12) + 0.45\widehat{\gamma}^{\star}(13).$$

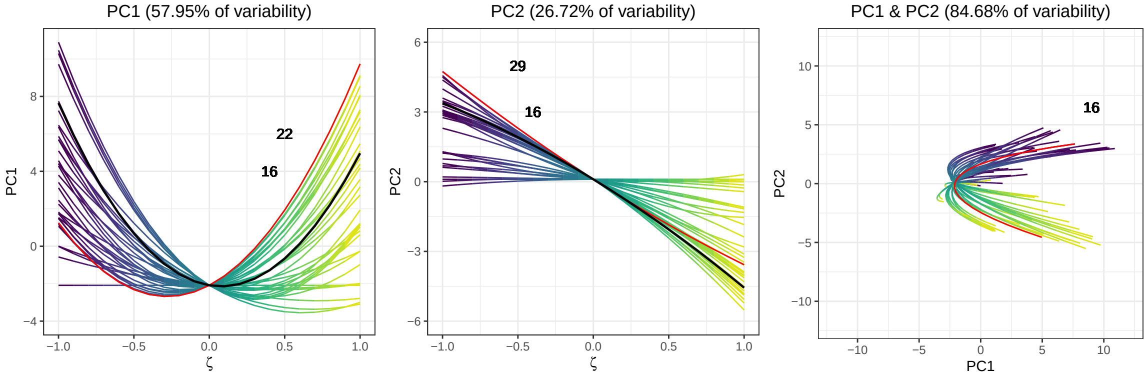
In Figure 20, we can see that PC1 had $79,52\%$ of the data variability, and the most influential function was at site $\#19$ ($\zeta > 0$) and for PC2 had 16.90%, and the most influential function was $\#9$ ($\zeta < 0$). These observations correspond to the values at location (-60.40, 46.67) and (-66.37, 45.98,), respectively. As seen for semivariogram and cross-semivariogram estimates, PC1 took information of all lags.

Figura 20 – From left to right: PC hair-plots for PC1 and PC2 with 79.52% and 16.90% of sample trace-semivariogram variability, respectively, and PC bihair-plot crossing PC1 and PC2 containing 96.43% of sample trace-semivariogram cumulative variability, considering spatial lag distance $h = \{1, 2, 3, 4\}$ and $\zeta \in [-1, 1]$ - maritimes data



Source: Elaborated by the author (2021).

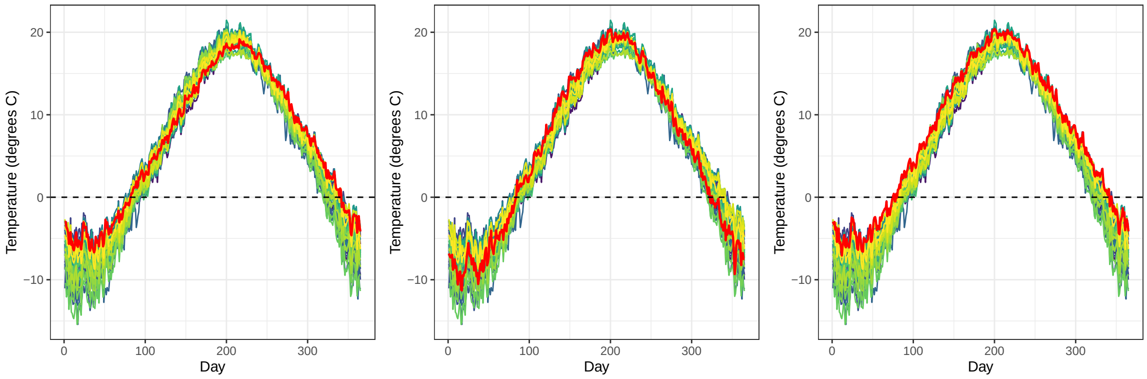
Figura 21 – From left to right: PC hair-plots for PC1 and PC2 with 57.88% and 26.72% of sample trace-semivariogram variability, respectively, and PC bihair-plot crossing PC1 and PC2 containing 84.68% of sample trace-semivariogram cumulative variability, considering spatial lag distance $h = 1, \ldots, 4$ and $\zeta \in (-1, 1)$ - maritimes data



Source: Elaborated by the author (2021).

In Figure 22, it can be seen that at the location $(-63.50, 44.63)$ of point $\#22$ the temperature becomes the highest at the end of the period, while that of point $\#29$ (at location $(-66.47, 45.85)$ stands out only when temperatures rise (middle of the period) and at $\#16$ (at location $(-64.85, 44.23)$ remains at highest for almost the entire period.

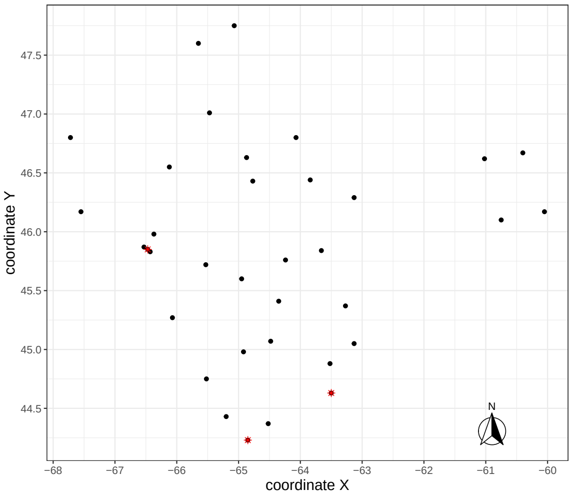
Figura 22 – From left to right: matplot highlighting in red the points corresponding to the observed location 22, 29 and 16, respectively- maritimes data



Source: Elaborated by the author (2021).

From the results, possible influential points were identified, being discarded the one found for $h = 1$: observation $\#16$, and those found relating the estimates of all lags: observation $\#22$ and $\#29$.

Figura 23 – Averages locations of daily temperature at 35 weather stations. The points in red correspond to possible influential points ($\#$16,$\#$22 and $\#$29), at locations $(-64.85, 44.23)$, $(-63.5, 44.63)$ and $(-66.47, 45.85)$, respectively - maritimes data.



Source: Elaborated by the author (2021).

6 SUPPLEMENTARY MATERIALS

Below is a list of routines and functions implemented in R, which will be made available after the article is published.

R code for semivariog.hair: R code for the command semivariogram hair (semivariog-hair.R). (R-code.zip);

R code for crosssemivariog.mm: R code for the command method-of-moments cross-semivariogram (crossvariogram-mm.R). (R-code.zip);

R code for crosssemivariog.mve: R code for the command minimum volume ellipsoid cross-semivariogram (crosssemivariog-mve.R). (R-code.zip);

R code for crosssemivariog.hair: R code for the command cross-semivariogram hair (cross-semivariog-hair.R). (R-code.zip);

R code for tracesemivariog.hair: R code for the command method-of-moments trace-semivariogram hair (trace-semivariog-hair.R). (R-code.zip);

R code for identify curve on hair-plot/bihair-plot: R code for command ggidentify (ggidentify.R);

Data: The pollution data are in Genton and Ruiz-Gazen (2010), the meuse data are in the R package sp and the maritimes data are in the R package geofd.

7 CONCLUSIONS

In this master's thesis, we adapted the hair-plot function for spatially dependent data. We built three types of graphs to identify influential observations: bihair-plot, PC hair-plot and PC bihair-plot. We generate such graphs for the pollution data considering the reflectance residual values for univariate case and obtain the influence on the semivariogram and perturbed values in the interval $[-40, 40]$. The most influential observation was $\#17$, which corresponds to the maximum value 40 at location $(2, 2)$ for positive perturbation values. We also identified influential observations for the meuse data, considering the variables zinc and lead, that is, for the bivariate case, and obtaining the influence on the cross-semivariogram for perturbation in $[-1, 1]$. For a positive perturbation, $\#54$ was the most influential observation of the data, with values equal to 1839 ppm of zinc and 654 ppm of lead, at location $(179973, 332255)$ in meters, also corresponding to the maximum values of the variables. As for functional data, we smoothed from the B-spline basis and we adapted the hair-plot for the trace-semivariogram estimator obtaining the functions of points $\#22, \#29$ and $\#16$ as possible influencing functions, the first being evaluated for a positive perturbation $\zeta > 0$ and the last two for a negative perturbation ($\zeta < 0$), at location (-66.37, 45.98), ($64.85, 44.23$) and ($63.50, 44.63$), respectively. Therefore, by applying the principal component analysis to values of spatial dependence estimators, it is possible to identify an influential observation and its location.

8 FUTURE WORKS

Considering that in the literature a graphical tool was presented to detect influential points in dependent data, and in particular, the spatial dependence is evaluated through the behavior of the estimated semivariance for different lags. The hair-plot presented by Genton and Ruiz-Gazen (2010) evaluates the influence through the additive perturbation for each lag, not taking information from the other lags. The main contribution presented in this master's thesis was to propose a methodology such that influential points are detected taking information from the spatial correlation present in the data. For its development, principal component analysis was used to construct the hair-plot and, in addition, the semi-variance obtained from the disturbed data was evaluated for paired lags, detecting influential points between two lags. For future work, the following items can be taken into account:

- Apply the methodology of functional data analysis smoothing non-parametric functions to these estimates and apply functional principal component analysis to obtain a curve for each observation and generating the functional hair-plot, since for each observation an estimation vector obtained from a vector of perturbations is evaluated;

- Evaluate the estimates obtained for only one additive perturbation and apply functional data analysis, adjusting non-parametric curves and generating the functional boxplot to detect possible outliers (see Genton and Sun (2014));

- Study other methods used to detect outliers in functional data to compare with the proposed tools (see Genton and Sun (2014));

- Investigate the local and asymptotic influence of estimators from real applications and simulations.

9 CONCLUSIONS

In this master's thesis, we adapted the hair-plot function for spatially dependent data. We built three types of graphs to identify influential observations: bihair-plot, PC hair-plot and PC bihair-plot. We generate such graphs for the pollution data considering the reflectance residual values for univariate case and obtain the influence on the semivariogram and perturbed values in the interval $[-40, 40]$. The most influential observation was $\#17$, which corresponds to the maximum value 40 at location $(2, 2)$ for positive perturbation values. We also identified influential observations for the meuse data, considering the variables zinc and lead, that is, for the bivariate case, and obtaining the influence on the cross-semivariogram for perturbation in $[-1, 1]$. For a positive perturbation, $\#54$ was the most influential observation of the data, with values equal to 1839 ppm of zinc and 654 ppm of lead, at location $(179973, 332255)$ in meters, also corresponding to the maximum values of the variables. As for functional data, we smoothed from the B-spline basis and we adapted the hair-plot for the trace-semivariogram estimator obtaining the functions of points $\#22, \#29$ and $\#16$ as possible influencing functions, the first being evaluated for a positive perturbation $\zeta > 0$ and the last two for a negative perturbation ($\zeta < 0$), at location (-66.37, 45.98), ($64.85, 44.23$) and ($63.50, 44.63$), respectively. Therefore, by applying the principal component analysis to values of spatial dependence estimators, it is possible to identify an influential observation and its location.

10 FUTURE WORKS

Considering that in the literature a graphical tool was presented to detect influential points in dependent data, and in particular, the spatial dependence is evaluated through the behavior of the estimated semivariance for different lags. The hair-plot presented by Genton and Ruiz-Gazen (2010) evaluates the influence through the additive perturbation for each lag, not taking information from the other lags. The main contribution presented in this master's thesis was to propose a methodology such that influential points are detected taking information from the spatial correlation present in the data. For its development, principal component analysis was used to construct the hair-plot and, in addition, the semi-variance obtained from the disturbed data was evaluated for paired lags, detecting influential points between two lags. For future work, the following items can be taken into account:

- Apply the methodology of functional data analysis smoothing non-parametric functions to these estimates and apply functional principal component analysis to obtain a curve for each observation and generating the functional hair-plot, since for each observation an estimation vector obtained from a vector of perturbations is evaluated;
- Evaluate the estimates obtained for only one additive perturbation and apply functional data analysis, adjusting non-parametric curves and generating the functional boxplot to detect possible outliers (see Genton and Sun (2014));
- Study other methods used to detect outliers in functional data to compare with the proposed tools (see Genton and Sun (2014));
- Investigate the local and asymptotic influence of estimators from real applications and simulations.

REFERENCES

AELST, S. V.; ROUSSEEUW, P. Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 1, n. 1, p. 71–82, 2009.

AZEVEDO, L.; PEREIRA, M. J.; RIBEIRO, M. C.; SOARES, A. Geostatistical covid-19 infection risk maps for portugal. *International Journal of Health Geographics*, BioMed Central, v. 19, n. 1, p. 1–8, 2020.

BABA, A. M.; MIDI, H.; ADAM, M. B.; RAHMAN, N. H. B. A. Detection of influential observations in spatial regression model based on outliers and bad leverage classification. Preprints, 2021.

BAI, J.; DEUTSCH, C. V. The pairwise relative variogram. *Geostatistics Lessons*, 2020.

BIVAND, R. S.; PEBESMA, E.; GOMEZ-RUBIO, V. *Applied spatial data analysis with R, Second edition*. Springer, NY, 2013. Available at: <https://asdar-book.org/>.

BORSSOI, J. A.; BASTIANI, F. D.; URIBE-OPAZO, M. A.; GALEA, M. Local influence of explanatory variables in gaussian spatial linear models. *The Chilean Journal of Statistics*, v. 2, n. 2, p. 29–38, 2011.

CASTRO, C. T. de. *Centro de Ciências Matemáticas e da Natureza Instituto de Matemática Departamento de Métodos Estatísticos*. Phd Thesis (PhD Thesis) — Universidade Federal do Rio de Janeiro, 2013.

CATTELL, R. B. The scree test for the number of factors. *Multivariate behavioral research*, Taylor & Francis, v. 1, n. 2, p. 245–276, 1966.

COOK, R. D. Detection of influential observation in linear regression. *Technometrics*, Taylor & Francis Group, v. 19, n. 1, p. 15–18, 1977.

CORTÉS-D, D. L.; CAMACHO-TAMAYO, J. H.; GIRALDO, R. Spatial prediction of soil penetration resistance using functional geostatistics. *Scientia Agricola*, SciELO Brasil, v. 73, p. 455–461, 2016.

CRESSIE, N. Geostatistics. *The American Statistician*, Taylor & Francis, v. 43, n. 4, p. 197–202, 1989.

CRESSIE, N. *Statistics for spatial data*. [S.l.]: John Wiley & Sons, 1993.

CRESSIE, N.; HAWKINS, D. M. Robust estimation of the variogram: I. *Journal of the international Association for Mathematical Geology*, Springer, v. 12, n. 2, p. 115–125, 1980.

De Bastiani, F.; CYSNEIROS, A. H. M. de A.; URIBE-OPAZO, M. A.; GALEA, M. Influence diagnostics in elliptical spatial linear models. *Test*, Springer, v. 24, n. 2, p. 322–340, 2015.

EVERITT, B.; HOTHORN, T. *An introduction to applied multivariate analysis with R*. [S.l.]: Springer Science & Business Media, 2011.

FERRATY, F.; VIEU, P. *Nonparametric functional data analysis: theory and practice*. [S.l.]: Springer Science & Business Media, 2006.

FOX, A. J. Outliers in time series. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 34, n. 3, p. 350–363, 1972.

GENTON, G. M.; SUN, Y. Functional data visualization. *John Wiley & Sons, Ltd*, Wiley StatsRef: Statistics Reference Online, 2020.

GENTON, M. G. Variogram fitting by generalized least squares using an explicit formula for the covariance structure. *Mathematical Geology*, Springer, v. 30, n. 4, p. 323–345, 1998.

GENTON, M. G.; RONCHETTI, E. Robust indirect inference. *Journal of the American Statistical Association*, Taylor & Francis, v. 98, n. 461, p. 67–76, 2003.

GENTON, M. G.; RUIZ-GAZEN, A. Visualizing influential observations in dependent data. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 19, n. 4, p. 808–825, 2010.

GENTON, M. G.; SUN, Y. Functional data visualization. *Wiley StatsRef: Statistics Reference Online*, Wiley Online Library, p. 1–11, 2014.

GIRALDO, R.; MATEU, J.; DELICADO, P. geofd: an r package for function-valued geostatistical prediction. *Revista Colombiana de Estadística*, Universidad Nacional de Colombia., v. 35, n. 3, p. 385–407, 2012.

JOLLIFFE, I. T. Discarding variables in a principal component analysis. i: Artificial data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 21, n. 2, p. 160–173, 1972.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, The Royal Society Publishing, v. 374, n. 2065, p. 20150202, 2016.

JONATHAN, R.; URIBE-OPAZO, M. A.; BASTIANI, F. d.; JOHANN, J. A. Técnicas para detecção de pontos influentes em variáveis contínuas regionalizadas. *Engenharia Agrícola*, SciELO Brasil, v. 36, p. 152–165, 2016.

LARK, R. Two robust estimators of the cross-variogram for multivariate geostatistical analysis of soil properties. *European Journal of Soil Science*, Wiley Online Library, v. 54, n. 1, p. 187–202, 2003.

MARTÍN, J. A. R.; ARIAS, M. L.; CORBÍ, J. M. G. Heavy metals contents in agricultural topsoils in the ebro basin (spain). application of the multivariate geoestatistical methods to study spatial variations. *Environmental pollution*, Elsevier, v. 144, n. 3, p. 1001–1012, 2006.

MARTIN, R. J. Leverage, influence and residuals in regression models when observations are correlated. *Communications in statistics-theory and methods*, Taylor & Francis, v. 21, n. 5, p. 1183–1212, 1992.

MATHERON, G. Principles of geostatistics. *Economic geology*, Society of Economic Geologists, v. 58, n. 8, p. 1246–1266, 1963.

PEBESMA, E. J.; BIVAND, R. S. Classes and methods for spatial data in R. *R News*, v. 5, n. 2, p. 9–13, November 2005. Available at: <https://CRAN.R-project.org/doc/Rnews/>.

RAMSAY, J. O.; DALZELL, C. Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 53, n. 3, p. 539–561, 1991.

RAMSAY, J. O.; SILVERMAN, B. W. *Functional Data Analysis: methods and case studies*. [S.l.]: Springer, 2005.

ROUSSEEUW, P. J. Least median of squares regression. *Journal of the American statistical association*, Taylor & Francis, v. 79, n. 388, p. 871–880, 1984.