Austrian Journal of Statistics
2026, Volume 55, Issue 3, 16–35.
http://www.ajs.or.at/

Visual Tools for Detecting Influential Observations in Bivariate Geostatistical Data

Jerfson B. N. Honório
Department of Statistics
Federal University of Pernambuco

Fernanda de Bastiani
Department of Statistics
Federal University of Pernambuco

Isabel S. D. de Oliveira
Department of Statistics
Federal University of Pernambuco

Manuel J. Galea Rojas
Facultad de Matematicas
Pontificia Universidad Católica de Chile

Abstract

This paper presents an extension of the hairplot method for detecting and visualizing influential observations in bivariate geostatistical models. To overcome the limitation of considering a single lag in semivariogram construction, we incorporate Andrews curves, allowing for a more comprehensive analysis. Additionally, we propose a novel approach that integrates boundary curves, providing a more rigorous methodology for detecting influential points. The effectiveness of the proposed methodology is assessed through simulation studies under different scenarios and disturbance levels and further demonstrated using a real soil dataset from southern Wisconsin. This application offers valuable insights into the impact of land management on carbon and nitrogen storage. By combining hairplots, cross-semivariograms, Andrews curves, and boundary curves, our approach enhances the diagnostic capabilities of spatial data analysis. This paper presents an extension of the hairplot method for detecting and visualizing influential observations in bivariate geostatistical models. To overcome the limitation of considering a single lag in semivariogram construction, we incorporate Andrews curves, allowing for a more comprehensive analysis. Additionally, we propose a novel approach that integrates boundary curves, providing a more rigorous methodology for detecting influential points. The effectiveness of the proposed methodology is assessed through simulation studies under different scenarios and disturbance levels and further demonstrated using a real soil dataset from southern Wisconsin. This application offers valuable insights into the impact of land management on carbon and nitrogen storage. By combining hairplots, cross-semivariograms, Andrews curves, and boundary curves, our approach enhances the diagnostic capabilities of spatial data analysis.

Keywords: Andrews curves, asymptotic influence, bivariate hairplot, boundary curves, cross semivariogram.

1. Introduction

There is a growing interest in data visualization tools in different areas of knowledge and also as support for identifying influential observations. In the following, we highlight some works in the literature that present these tools in the context of the topic of this article. Genton and Ruiz-Gazen (2010) proposed a tool to visualize influential observations in the context of dependent data based on the study of perturbations in the data and the effect on the estimators $\widehat{\theta}(\mathbf{Z})$ of a parameter $\theta(\mathbf{Z})$, and introduced the graph known as the *hairplot* to detect and analyze influential observations. According to the authors, an observation is influential if it causes a radical change in the values of the respective estimates. Sun and Genton (2011) presented boxplots for functional data and Yao, Dai, and Genton (2020) proposes two informative exploratory tools: the functional trajectory boxplot and the modified simplicial band depth versus wiggliness of directional outlyingness plot, to visualize the centrality of trajectory functional data. More recently, Ojo, Anta, Genton, and Lillo (2023) introduced definitions and properties of fast indices for unsupervised outlier detection, used to identify outliers in functional data. Additionally, Jiménez-Varón, Harrou, and Sun (2024) presented an innovative approach for univariate and multivariate functional outlier detection, utilizing pointwise depth distribution and correlations between pairwise depths, with applications in solar energy data. On the other hand, Alcacer and Epifanio (2024) introduced an innovative methodology for anomaly detection in clustered functional data, extending the KNN technique to functional data contexts and demonstrating superiority compared to state-of-the-art methods. Furthermore, Qu, Dai, Euan, Sun, and Genton (2025) provided a comprehensive review of recent procedures for exploratory functional data analysis (EFDA), including visualization, outlier detection, and clustering techniques, with implementations available.

These recent developments highlight the importance of combining graphical tools and robust statistical diagnostics, especially when dealing with complex data structures. Our work builds upon this direction by proposing a visualization-based perturbation method for detecting influential observations in spatially dependent multivariate data.

Identifying influential points and outliers in spatial data represents a fundamental aspect of exploratory and diagnostic spatial analysis. In the case of dependent observations, as is the case with spatial data, the influence function is defined in terms of the joint distribution of the data. In this context, the evaluation of the influence function is conducted using a method that involves additive perturbation. In their study, Genton and Ruiz-Gazen (2010) proposed a tool for visualizing potentially influential observations. The tool is based on the analysis of the impact of data perturbation on the estimators of a parameter, with a particular focus on the estimator of the parameter, which is represented by the function $\hat{\theta}(\mathbf{z})$. The authors proposed the hairplot as a methodology for the identification and examination of points of influence. In order to develop this tool, the researchers defined an empirical influence based on the additive perturbation $\mathbf{z}[i, \zeta]$, considering a perturbation value $\zeta \in \mathbb{R}$. This provided further insight into the behaviour of the estimators, specifically how they respond to perturbations in the data. Two influential measures were proposed: the local and asymptotic influence of the i-th observation. An increase in the absolute value of the initial measurement indicates a greater influence on the observation, whereas the second measurement indicates the influence on the estimator value when ζ at a relatively large value. In the paper, the variable $\mathbf{z}(\mathbf{s})$ represents a spatial process, and the function $\hat{\theta}(\cdot)$ denotes the cross semivariogram of the method of moments.

The objective of this paper goes beyond proposing an extension of the hairplot to the bivariate geostatistical context. The proposed approach incorporates boundary curves into the analysis, providing a more automated method for outlier detection and allowing a deeper understanding of the influence of specific observations. It also addresses the limitation of considering only a single lag. Now, in both univariate and bivariate geostatistical settings, the hairplot can include all lags of the semivariogram, improving its applicability and quality.

2. Geostatistics

The semivariogram is a statistical measure that assesses the degree of correlation between two samples that are separated by a distance vector, denoted by $\mathbf{h}$. The regionalized variable $Z(\mathbf{s})$ is associated with sample locations $\mathbf{s}$ and $\mathbf{s}+\mathbf{h}$, which are expressed as vectors in $\mathbb{R}^d$. In this context, the notation $Z(\mathbf{s})$ is used to represent a specific value of the regionalized variable Z. The subset of real numbers in a d-dimensional space represents an intrinsically stationary process defined within a domain $\mathcal{D}$. As defined by Cressie (2015), the semivariogram, denoted by $\gamma(\mathbf{h})$, can be expressed as follows:

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

There are multiple methodologies for estimating the semivariogram. The method of moments, as outlined by Matheron (1963), is a commonly utilized approach in the scientific literature. However, this estimator may be susceptible to the influence of outliers. As an alternative, other more robust estimators were proposed, including those of Cressie and Hawkins (1980), Genton (1998), Babakhani and Deutsch (2012), among others. In this article, we have elected to employ exclusively the method of moments estimator.

The semivariogram formulation proposed by Matheron (1963) is as follows:

$$\widehat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\left[Z\left(s_i+\mathbf{h}\right) - Z\left(s_i\right)\right]^2 = \frac{1}{2}\mathbf{z}^\top \mathbf{A}(\mathbf{h})\mathbf{z},$$

where $\hat{\gamma}(\mathbf{h})$ represents the semivariogram estimate, $Z\left(s_i\right)$ denotes the value of the variable Z at position s_i, and $Z\left(s_i+\mathbf{h}\right)$ is the value of the variable Z at position $s_i+\mathbf{h}$. The term $N(\mathbf{h})$ denotes the number of pairs that are separated by a given distance $\mathbf{h}$. In the case of an irregularly sampled data grid, the value of $N(\mathbf{h})$ can be expressed as a set of ordered pairs, where each pair consists of two distinct grid points, s_i and s_j with a distance between them given by $_i - s_j$. This is represented by $T(\mathbf{h})$, where $T(\mathbf{h})$ is a subset of R^d and encompasses $\mathbf{h}$, as cited in Cressie (2015).

Furthermore, the expression $\mathbf{z}^\top\mathbf{A}(\mathbf{h})\mathbf{z}$ denotes the quadratic form proposed by Genton (1998), where $\mathbf{z} = (Z_1, Z_2, \ldots, Z_n)^\top$, and $\mathbf{A}(\mathbf{h})$ is the spatial design matrix for the data at lag $\mathbf{h}$, as outlined by the same author.

2.1. Cross semivariogram

The cross semivariogram is a statistical tool used to assess the degree of association between two regionalized variables, $Z_u = \left(Z_u\left(s_1\right), \ldots, Z_u\left(s_n\right)\right)^\top$ and $Z_v = \left(Z_v\left(s_1\right), \ldots, Z_v\left(s_n\right)\right)^\top$. It is assumed that both are intrinsically stationary processes defined in a domain $\mathcal{D}$, where $\left\{Z_u(\mathbf{s}), Z_v(\mathbf{s}) : \mathbf{s} \in \mathcal{D} \subset \mathbb{R}^d\right\}$. This measure evaluates the association between these variables and is defined as:

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

where $(\mathbf{s})$ and $(\mathbf{s}+\mathbf{h})$ are sample locations, and $\mathbf{h}$ represents the euclidean distance between these locations.

According to Lark (2003), the cross semivariogram obtained from the method of moments is defined by

$$\widehat{\gamma}_{u,v}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\left[Z_u\left(s_i+\mathbf{h}\right) - Z_u\left(s_i\right)\right]\left[Z_v\left(s_i+\mathbf{h}\right) - Z_v\left(s_i\right)\right] = \frac{1}{2}\mathbf{z}_u^\top\mathbf{A}(\mathbf{h})\mathbf{z}_v,$$

where $\frac{1}{2}\mathbf{z}_u^\top\mathbf{A}(\mathbf{h})\mathbf{z}_v$ represents the quadratic form of the cross semivariogram estimator; $\widehat{\gamma}_{u,v(\mathbf{h})}$ is the value of the cross semivariogram estimate; $Z_u\left(s_i\right)$ and $Z_v\left(s_i\right)$ are the values of the

variables Z_u and Z_v, respectively, at position s_i; $Z_u(s_i + \mathbf{h})$ and $Z_v(s_i + \mathbf{h})$ are the values of the variables Z_u and Z_v, respectively, at position $s_i + \mathbf{h}$; $N(\mathbf{h}) = \{(s_i, s_j) : ||s_i - s_j|| = \mathbf{h}\}$ is the number of pairs separated by a given distance $\mathbf{h}$.

For an irregularly sampled data grid, $N(\mathbf{h}) = \{(s_i, s_j) : s_i - s_j \in T(\mathbf{h})\}$, where $T(\mathbf{h}) \subset \mathbb{R}^d$ around $\mathbf{h}$ (Cressie 2015). In this case, when evaluating the range and threshold, the focus is on studying the maximum distance of spatial dependence and the approximation of the covariance between the two variables, respectively. The Cauchy-Schwartz relation, given by: $|\gamma_{u,v}| = \sqrt{\gamma_u \gamma_v}$; ensures that all distances h considered in a cokriging process are guaranteed.

2.2. Additive perturbation

In order to gain a more detailed understanding of the estimator's behaviour when adding low perturbation values, Genton and Ruiz-Gazen (2010) introduced an additive perturbation framework. Let us consider the data vector $\mathbf{z} = (Z_1, Z_2, \ldots, Z_n)^\top$ and an additive perturbation of $\mathbf{z}$, given by:

$$\mathbf{z}[i, \zeta] = \mathbf{z} + \zeta \mathbf{e}_i,$$

where $\mathbf{e}_i$ is a matrix of zeros with the i-th element equal to 1, and $\zeta \in \mathbb{R}$ represents the amount of perturbation. For $\zeta = 0$, the value $\mathbf{z}$ is preserved for all observations. This framework allows analysis of estimator behavior at different levels of perturbation, providing valuable insights into the sensitivity of the estimator to variations in the data.

2.3. Local and asymptotic influence

Let $\widehat{\theta}(\cdot)$ be the estimator of θ, and $\widehat{\theta}(\mathbf{z}[i, \zeta])$ be the perturbed estimator of the data for all $i = 1, \ldots, n$ and ζ. Genton and Ruiz-Gazen (2010) showed that by plotting each of these estimators, it is possible to visualize the effect of influential observations. The "hairplot" is therefore a version of the empirical influence function ((Hampel, Ronchetti, Rousseeuw, and Stahel 1986)) with replacement. Furthermore, they proposed two influence measures: the local influence function and the asymptotic influence function. The local influence of the i-th observation is defined in Equation (1):

$$\begin{aligned}
\tau_i(\widehat{\theta}, \mathbf{z}) &= \frac{\partial}{\partial \zeta} \widehat{\theta}(\mathbf{z}[i, \zeta]) \Big|_{\zeta=0} \\
&= \frac{\partial}{\partial \zeta} \widehat{\gamma}_{u,v}(\mathbf{z}[i, \zeta]) \Big|_{\zeta=0} \\
&= \frac{\partial}{\partial \zeta} \left[\mathbf{z}_{\mathrm{u}}^\top \mathbf{A}(\mathbf{h}) \mathbf{z}_{\mathrm{v}} + \left(\mathbf{z}_{\mathrm{u}}^\top \mathbf{A}(\mathbf{h}) \mathbf{e}_i + \mathbf{e}_i^\top \mathbf{A}(\mathbf{h}) \mathbf{z}_{\mathrm{v}} \right) \zeta + \left(\mathbf{e}_i^\top \mathbf{A}(\mathbf{h}) \mathbf{e}_i \right) \zeta^2 \right] \Big|_{\zeta=0} \\
&= \mathbf{z}_{\mathrm{u}}^\top \mathbf{A}(\mathbf{h}) \mathbf{e}_i + \mathbf{e}_i^\top \mathbf{A}(\mathbf{h}) \mathbf{z}_{\mathrm{v}}.
\end{aligned} \tag{1}$$

Therefore, the largest absolute value of $\tau_i(\cdot)$ represents the observation with the greatest influence. Conversely, the asymptotic influence of the i-th observation implies the influence of the estimate $\hat{\theta}(\mathbf{z})$ when there is a high perturbation value for the i-th observation given in Equation (2):

$$\begin{aligned}
\nu_i(\widehat{\theta}, \mathbf{z}) &= \lim_{\zeta \to \infty} \widehat{\theta}(\mathbf{z}[i, \zeta]) \\
&= \lim_{\zeta \to \infty} \widehat{\gamma}_{u,v}(\mathbf{z}[i, \zeta]) \\
&= \lim_{\zeta \to \infty} \left[\mathbf{z}_{\mathrm{u}}^\top \mathbf{A}(\mathbf{h}) \mathbf{z}_{\mathrm{v}} + \left(\mathbf{z}_{\mathrm{u}}^\top \mathbf{A}(\mathbf{h}) \mathbf{e}_i + \mathbf{e}_i^\top \mathbf{A}(\mathbf{h}) \mathbf{z}_{\mathrm{v}} \right) \zeta + \left(\mathbf{e}_i^\top \mathbf{A}(\mathbf{h}) \mathbf{e}_i \right) \zeta^2 \right] \\
&= \infty.
\end{aligned} \tag{2}$$

In order to illustrate the hairplot in the bivariate geostatistical context, it is necessary to consider the representation provided below. At each iteration, a single observation is subjected to a random perturbation, after which the semivariance is calculated.

$$\mathbf{z}[1,\zeta] = \begin{pmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \\ \vdots & \vdots \\ Z_{n1} & Z_{n2} \end{pmatrix} + \zeta \begin{pmatrix} 1 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} Z_{11}+\zeta & Z_{12}+\zeta \\ Z_{21} & Z_{22} \\ \vdots & \vdots \\ Z_{n1} & Z_{n2} \end{pmatrix} \longrightarrow \widehat{\gamma}(h)^{(1,\zeta)}$$

$$\vdots$$

$$\mathbf{z}[n,\zeta] = \begin{pmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \\ \vdots & \vdots \\ Z_{n1} & Z_{n2} \end{pmatrix} + \zeta \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \\ \vdots & \vdots \\ Z_{n1}+\zeta & Z_{n2}+\zeta \end{pmatrix} \longrightarrow \widehat{\gamma}(h)^{(n,\zeta)}$$

where $\widehat{\gamma}(h)^{(n,\zeta)}$ represents the semivariance at lag h, with the observation n perturbed by ζ. As we have a new variable h, we now deal with two random variables, h and ζ, transforming vectors into matrices. This is crucial for the introduction of matrix notation and for future work with Andrews plot, as we will discuss later Moustafa (2011). Thus, the semivariogram matrix is given by:

$$\mathbf{\Gamma}(\mathbf{h},\zeta) = \begin{bmatrix} \widehat{\gamma}^{(1,\zeta)}(h_1) & \cdots & \widehat{\gamma}^{(1,\zeta)}(h_p) \\ \vdots & \ddots & \vdots \\ \widehat{\gamma}^{(n,\zeta)}(h_1) & \cdots & \widehat{\gamma}^{(n,\zeta)}(h_p) \end{bmatrix}. \tag{3}$$

Therefore, the hairplot is constructed considering the ζ perturbations versus the semivariance vector, considering h as fixed. It should be noted that this approach is subject to a limitation when only a fixed lag is considered, with lags up to a distance of 50% of the maximum distance between observations being investigated ((Cressie 2015)).

3. Boundary curves for hairplots

This section introduces an innovative approach to constructing boundary curves in hairplot using boxplots to identify potentially influential observations. The concept is derived from the pocket plot methodology ((Cressie 2015)), wherein the number of rows is fixed and a boxplot is generated for each row. This approach is then applied to the context of perturbations.

To represent the perturbations applied to the observations, we use a perturbation matrix $\mathbf{\Gamma}(h_1,\zeta)$, where each column corresponds to a perturbation value ζ applied to all observations for a given lag h_1. The matrix is defined as follows:

$$\mathbf{\Gamma}(h_1,\zeta) = \begin{bmatrix} \widehat{\gamma}^{(1,\zeta_1)}(h_1) & \cdots & \widehat{\gamma}^{(1,\zeta_r)}(h_1) \\ \vdots & \ddots & \vdots \\ \widehat{\gamma}^{(n,\zeta_1)}(h_1) & \cdots & \widehat{\gamma}^{(n,\zeta_r)}(h_1) \end{bmatrix}.$$

In this context, the notation $\widehat{\gamma}^{(i,\zeta_j)}(h_1)$ represents the estimated semivariogram value for the i-th observation, with the perturbation ζ_j at lag h_1. Each row represents an observation, and each column represents a perturbation applied to the observations.

For each column of the matrix, a boxplot is computed in order to identify potentially influential observations. We use the traditional rule: an observation is considered influential if the value is above $(Q3 + 1.5 \times \text{IQR})$ or below $(Q1 - 1.5 \times \text{IQR})$, where $Q3$ means third quartile, $Q1$ first quartile and IQR is the interquartile distance. This process is repeated for each perturbation level ζ, forming boundary curves for each observation across the different perturbations.

This approach addresses one of the limitations of traditional hairplots, which is the difficulty in determining which observations are truly influential. The calculation of boundary curves calculated from the matrix $\mathbf{\Gamma}(h_1, \zeta)$ provides a clear and objective method for identifying influences with greater certainty across different perturbation levels.

This methodology offers a comprehensive and dependable perspective on the impact of observations, thereby facilitating a more accurate analysis of spatial data.

Algorithm 1 outlines the systematic steps for generating the boundary curves from the perturbation matrix.

Algorithm 1 Construction of Boundary Curves for hairplots

// – **Input: Perturbation matrix** $\Gamma(h, \zeta)$ **for a fixed lag** h –
For each perturbation level ζ_j in $\{\zeta_1, \zeta_2, ..., \zeta_r\}$:
  Extract column j from matrix $\Gamma(h, \zeta)$ (values for all n observations)
  Calculate the first quartile $Q1$
  Calculate the third quartile $Q3$
  Compute the interquartile range: $IQR_j = Q_3 - Q_1$

  // – **Define the fences (Tukey's rule)** –
  Set Upper Boundary at ζ_j: $UB_j = Q3 + 1.5 \times IQR_j$
  Set Lower Boundary at ζ_j: $LB_j = Q1 - 1.5 \times IQR_j$
  Store coordinates (UB_j, ζ_j) and (LB_j, ζ_j)
End For

// – **Visualization** –
Connect all points (UB_j, ζ_j) to form the Upper Boundary Curve
Connect all points (LB_j, ζ_j) to form the Lower Boundary Curve
Plot curves alongside the individual hairplots $\hat{\gamma}^{(i,\zeta)}(h)$

4. Andrews plot for identifying influential observations

In the context of multivariate statistical analysis, it is of paramount importance to effectively handle data represented by a random vector of dimension $(p > 1)$. A sample of size n from a random vector $\mathbf{z} = (Z_1, \ldots, Z_p)^\top$, can be used to form the data matrix $\mathbf{Z}$ given by

$$\mathbf{Z} = \begin{pmatrix} Z_{11} & Z_{12} & \ldots & Z_{1p} \\ Z_{21} & Z_{22} & \ldots & Z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \ldots & Z_{np} \end{pmatrix}. \tag{4}$$

The Andrews plot enables the representation of each multivariate data point by a curve, employing Fourier-based interpolation wherein the coefficients correspond to the components of the observation, as evidenced in Moustafa (2011). These curves, defined as linear combinations of the components of the data matrix (Eq. 4), are particularly useful for describing data structure, identifying outliers, and clustering observations ((Andrews 1972)). For each index $i = 1, \ldots, n$, the i-th Andrews curve is given by

$$X_i(t) = \frac{1}{\sqrt{2\pi}} Z_{i1} + \sin(t) Z_{i2} + \cos(t) Z_{i3} + \sin(2t) Z_{i4} + \cdots,$$

i.e.,

$$X_i(t) = \begin{cases} \dfrac{Z_{i,1}}{\sqrt{2}} + Z_{i,2}\sin(t) + Z_{i,3}\cos(t) + \cdots \\ \ldots + Z_{i,p-1}\sin\left(\dfrac{p-1}{2}t\right) + Z_{i,p}\cos\left(\dfrac{p-1}{2}t\right) & \text{for } p \text{ odd} \\ \dfrac{Z_{i,1}}{\sqrt{2}} + Z_{i,2}\sin(t) + Z_{i,3}\cos(t) + \cdots + Z_{i,p}\sin\left(\dfrac{p}{2}t\right) & \text{for } p \text{ pair,} \end{cases} \tag{5}$$

where $t \in [-\pi, \pi]$. The order of variables is an important factor in determining the shape of the curve, as variables introduced later contribute less to the overall shape. It is therefore common practice to order the variables in accordance with their informative contribution, frequently utilising principal component analysis.

Andrews curves, as defined in Equation (5), transform the multivariate sample into a sample of a functional random field $\left\{X_t(s), s \in D \subset \mathbb{R}^d, t \in [-\pi, \pi] \subset \mathbb{R}\right\}$, with the transformation:

$$X_t(s_i) = \sum_{k=1}^{p} Z_k(s_i)\,\phi_k(t),$$

where $\phi_k(t)$ represents the k-th coefficient of the Fourier series.

5. Andrews plot for detecting influential observations

The traditional hairplot is limited to visualizing influence at a single, fixed lag h. However, a comprehensive influence analysis requires assessing an observation's impact across a range of lags $\{h_1, h_2, ..., h_p\}$. For any given observation, this generates a vector of semivariance estimates, turning the diagnostic challenge into a multivariate problem. To address this, we propose the use of Andrews curves, a powerful technique for visualizing multivariate data. Each observation's vector of semivariance estimates across all lags is mapped to a unique curve, allowing for the simultaneous visual assessment of its overall behavior, the identification of clusters, and the detection of outliers that represent potentially influential points.

By employing Andrews plot, we introduce two new approaches. In the first approach, we fix a specific value of ζ, using the matrix $\mathbf{\Gamma}(\mathbf{h}, \zeta)$ from Equation (3), where each column represents a variable that depends exclusively on the lag h. This allows us to identify potentially influential observations by considering all lags for a given ζ. This solution addresses the limitation of the traditional hairplot, which allows visualization of influence for only a specific lag.

In the second approach, we introduce a new perturbation matrix that encompasses all values of ζ. The first approach faces a similar limitation to the traditional hairplot, as it only visualizes a single value of ζ. To overcome this, we propose a perturbation matrix that covers all lags and all values of ζ, defined as:

$$\mathbf{\Gamma}_{\top}(\mathbf{h}, \zeta) = \begin{bmatrix} \widehat{\gamma}^{(1,\zeta_1)}(h_1) & \cdots & \widehat{\gamma}^{(1,\zeta_1)}(h_p) \\ \vdots & \ddots & \vdots \\ \widehat{\gamma}^{(n,\zeta_1)}(h_1) & \cdots & \widehat{\gamma}^{(n,\zeta_1)}(h_p) \\ \widehat{\gamma}^{(1,\zeta_2)}(h_1) & \cdots & \widehat{\gamma}^{(1,\zeta_2)}(h_p) \\ \vdots & \ddots & \vdots \\ \widehat{\gamma}^{(n,\zeta_r)}(h_1) & \cdots & \widehat{\gamma}^{(n,\zeta_r)}(h_p) \end{bmatrix}. \tag{6}$$

The proposed matrix in Equation (6) aims to capture as much information as possible from the data, considering all lags and values of ζ, providing a comprehensive view of potential influences within the data.

In order to ensure full reproducibility and to make the proposed methodology transparent and accessible, we have included the complete pseudocode for the simulation and analysis directly in the main body of the paper.

Algorithm 2 details the steps involved in the data generation and contamination process, covering both scenarios considered in the simulation study.

Algorithm 2 hairplot and Andrews Curves Generation

// – **Input: A spatial dataset** –
For each lag h in $\{1, 2, ..., p\}$:
  For each perturbation level ζ:
    For each observation i in $\{1, ..., n\}$:
      Create a temporary copy of the dataset
      Apply an additive perturbation ζ to observation i
      Calculate the cross-semivariogram for lag h
      Store the resulting semivariance value $\gamma(h, i, \zeta)$
    End For
  End For
End For

// – **Visualization** –
For a fixed perturbation ζ (Andrews plot across lags):
  For each observation i:
    Create a vector $V_i = [\gamma(1, i, \zeta), \gamma(2, i, \zeta), ..., \gamma(p, i, \zeta)]$
    Generate and plot the Andrews curve for vector V_i.
  End For

6. Analysis and results

6.1. Simulation study design

To assess the performance of the proposed visual tools, we conducted a simulation study using the R programming language. The artificial dataset was generated by simulating a bivariate Gaussian Random Field (GRF) on a 9×9 regular spatial grid. To address the requirement for explicit spatial correlation, the process was modeled using an Exponential semivariogram with a range parameter of 5 units and a nugget effect of 0.5. This ensures that the empirical structure analyzed reflects a true geostatistical model rather than independent realizations arranged on a grid. The bivariate distribution was parameterized with a mean vector $\mu = [3, 3]^\top$ and a coregionalization matrix Σ:

$$\Sigma = \begin{pmatrix} 2.0 & 0.5 \\ 0.5 & 2.0 \end{pmatrix}.$$

Two contamination scenarios were created to simulate influential observations. In both scenarios, observation #21 was altered. In Scenario 1, only observation #21 of the first variable was perturbed. In Scenario 2, observation #21 for both variables were perturbed. The perturbation was applied additively at different levels, using the standard deviation of the original simulated data.

To analyze the behavior of the bivariate hairplot and its modifications using the Andrews plot, we generated artificial data from a bivariate normal distribution on a 9×9 grid, considering two different scenarios. In both cases, observation #21 was perturbed. In the first scenario, the perturbation affected only one of the variables, while in the second scenario, both variables were perturbed. These perturbations were applied at three different levels, corresponding to 1 and 2 times the standard deviation.

Algorithm 3, in turn, presents the procedures for computing the perturbed semivariograms and generating the visualizations based on hairplots and Andrews curves.

Algorithm 3 Simulation and Contamination Procedure

// – **Initialization** –

Set random seed to 21

Define mean vector $\mu = [3, 3]^\top$

Define covariance matrix $\Sigma = \begin{pmatrix} 2.0 & 0.5 \\ 0.5 & 2.0 \end{pmatrix}$

// – **Data Generation** –

Generate $n = 81$ samples from a Bivariate Gaussian Random Field (GRF) using an Exponential variogram (Range = 5, Nugget = 0.5)

Arrange data on a 9x9 regular grid

// – **Data Contamination** –

For Scenario 1 (perturb one variable):

  Identify observation #21

  Replace its value Z_1 with $Z_1 + \text{sd}(\text{all } Z_1)$

For Scenario 2 (perturb both variables):

  Identify observation #21

  Replace its value Z_1 with $Z_1 + \text{sd}(\text{all } Z_1)$

  Replace its value Z_2 with $Z_2 + \text{sd}(\text{all } Z_2)$

6.2. Artificial data

Scenario 1

In the initial scenario, a perturbation is applied to data from a bivariate normal distribution on a 9×9 grid, with a particular emphasis on observation #21. In this scenario, the perturbation affects a single variable in observation #21. Three distinct levels of perturbations are considered, as outlined below:

- $Z_1^{21}(s) = Z_1^{21}(s) + sd(Z^1(s))$,
- $Z_1^{21}(s) = Z_1^{21}(s) + 2sd(Z^1(s))$.

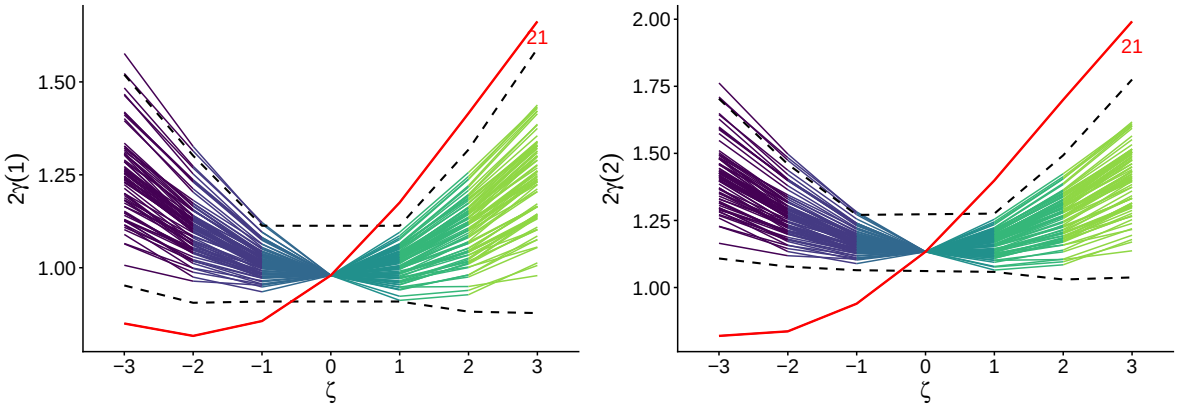
After the perturbations, hairplots were generated, as shown in Figure 1, using cross semivariance. Note that the identification of influential points is considerably enhanced when the hairplot and the Cross Semivariogram are employed in conjunction with one another in the analytical process. The anomalous observation is correctly identified, exhibiting a notable divergence from the other observations.

When the hairplot is used in conjunction with the Cross Semivariogram, there is a limitation in the ability to visualize individual lags. To address this limitation, we propose an innovative solution: the incorporation of the Andrews plot. This approach enables the visualization of all lags in a single-curve format, facilitating a comprehensive analysis that encompasses all potentially influential observations across all lags, thereby enhancing the ability to identify influential points.

In the context of Andrews's plot, two approaches are at our disposal. In the first approach, we fix the value of ζ. We will explore six different values for the standard deviations of $\zeta(-3, -2, -1, 1, 2, 3)$, which will result in six different matrices of $\mathbf{\Gamma}$, each associated with a particular value of ζ.

Therefore, given the availability of the $\mathbf{\Gamma}$ matrices, it is possible to generate an Andrews plot for each of these matrices. Figure 2 presents a hairplot based on Andrews curves for a fixed parameter, ζ.

Observation #21 exerts a notable influence across all lags, distinguishing itself from the others. When examining the revised matrix, designated as 6, we can construct new Andrews curves



(a) $h = 1$ (b) $h = 2$

Figure 1: Hairplot of a perturbed variable considering the 1 and 2 lags ($h = 1, 2$) in the cross semivariogram with $\zeta(-3, -2, -1, 0, 1, 2, 3)$

encompassing the entirety of the data captured by the perturbations and lags. Consequently, the hairplot, illustrated in Figure 3, incorporates all information pertaining to all lags and perturbation levels.

Figure 3 highlights observation #21, identified as the perturbed observation. Consequently, the Andrews curves, derived from matrix (6), demonstrate their effectiveness as a tool for visualizing potentially influential observations across all lags and perturbation levels. This approach is particularly effective in scenarios where only a single variable is perturbed."

Scenario 2

In the second scenario, the perturbation affected both variables in observation #21. The same three levels of perturbation as in scenario 1 were considered.

It is observed that all hairplots, presented in Figure 4, are extremely similar to those in scenario 1, with only small differences in the semivariance values.

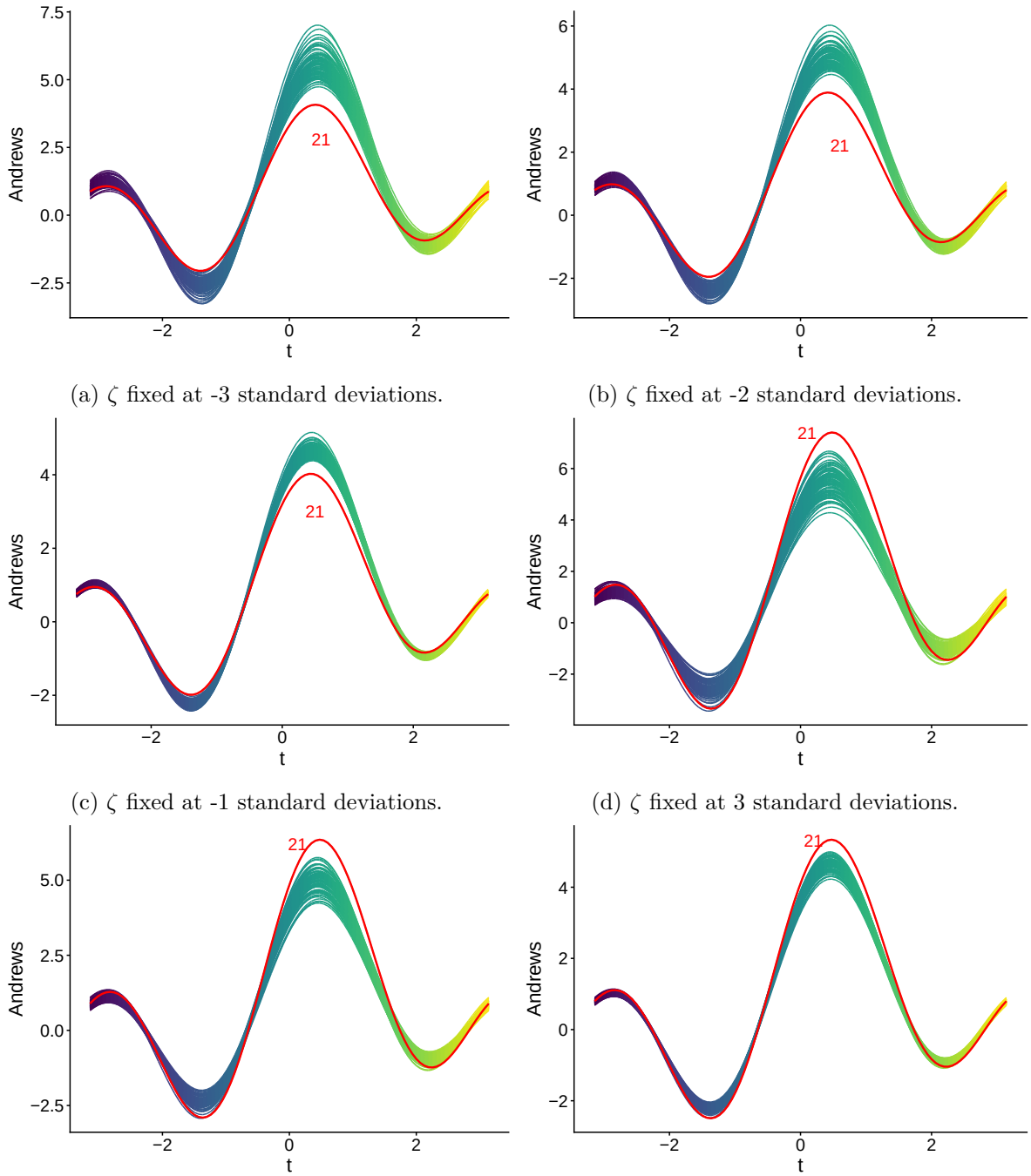
When employing Andrews Curves, we again obtain results very similar to those in scenario 1, as shown in Figures 5 and Figure 6.

It was determined that, in the proposed scenarios, the outcomes remain practically identical regardless of whether an influential observation is present in a single variable or in both variables. This finding highlights the quality of the analyses, as the nature of the influence exerted by the observations does not significantly affect the overall results. Consequently, these results underscore the credibility of the proposed method and the reliability of its conclusions, demonstrating that the presence or absence of influence on individual variables has minimal impact on the overall response.

6.3. Carbon and nitrogen data

The database under analysis originates from a study conducted in southern Wisconsin with the objective of understanding the impacts of diversified land management on soil carbon and nitrogen storage. This long-term study, conducted by researchers affiliated with SAGE, includes data collected from more than 125 distinct locations across a range of land types, including conventional farmland, conservation areas, restored prairie ecosystems, and grasslands.

The state of Wisconsin is distinguished by a diverse topography, encompassing a mosaic of agricultural lands, forested areas, grasslands, bodies of water, and urban settlements. The conversion of prairies and forests to farmland in the nineteenth century resulted in a notable depletion of organic matter in the soil, accompanied by the release of carbon dioxide (CO_2) into the atmosphere. It is therefore essential to identify potentially influential observations



(a) ζ fixed at -3 standard deviations.

(b) ζ fixed at -2 standard deviations.

(c) ζ fixed at -1 standard deviations.

(d) ζ fixed at 3 standard deviations.

(e) ζ fixed at 2 standard deviations.

(f) ζ fixed at 1 standard deviations.

Figure 2: Hairplot based on Andrews curves for a fixed parameter, ζ

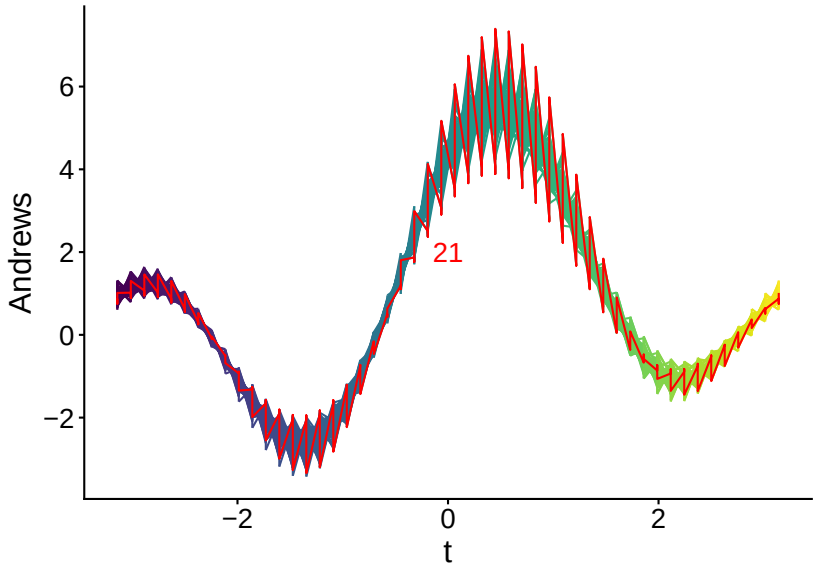
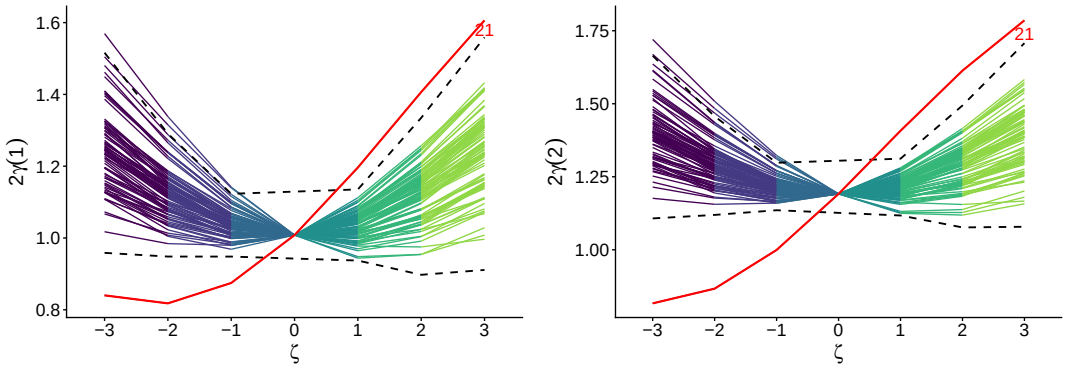
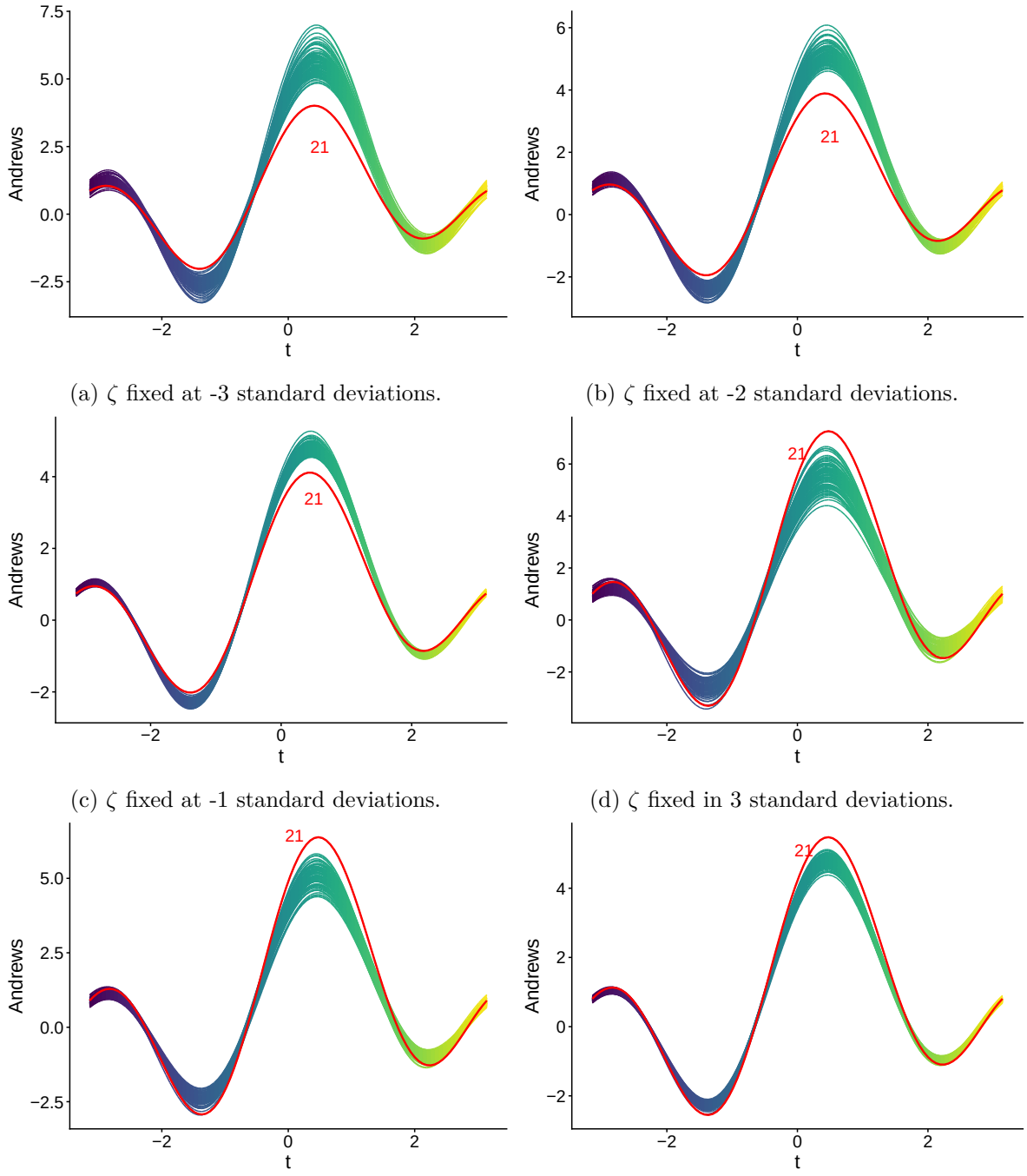


Figure 3: Hairplot based on Andrews curves for all ζ



(a) $h = 1$ (b) $h = 2$

Figure 4: Hairplot of a perturbed variable considering the 1 and 2 lags ($h = 1, 2$) in the cross semivariogram with $\zeta(-3, -2, -1, 0, 1, 2, 3)$



(a) ζ fixed at -3 standard deviations.

(b) ζ fixed at -2 standard deviations.

(c) ζ fixed at -1 standard deviations.

(d) ζ fixed in 3 standard deviations.

(e) ζ fixed in 2 standard deviations.

(f) ζ fixed at 1 standard deviations.

Figure 5: Hairplot based on Andrews curves for artificial data

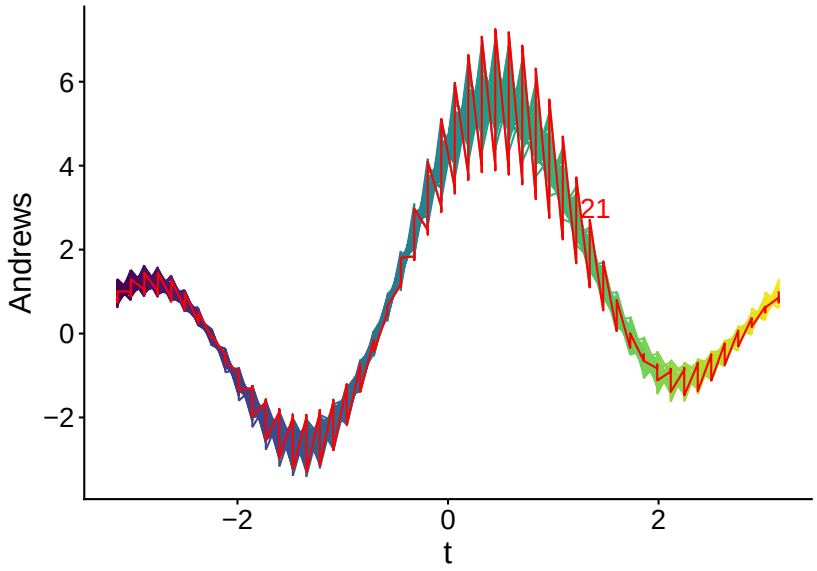


Figure 6: Hairplot based on Andrews curves for all ζ for the artificial data

in this context, as this allows for the assessment of the agricultural practices and areas that contribute most to soil degradation and carbon emissions.

The database contains information on Total Soil Organic Carbon (SOC) and Total Soil Organic Nitrogen (SON) in the 0-25 cm soil layer, expressed in kilograms per square meter. Additionally, the geographical coordinates of each sampled location were recorded, with adjustments for the UTM system to ensure precision in spatial analyses. Other collected variables include soil bulk density, land-use history, and the diversity of management practices at each site.

Data analysis will employ the hairplot alongside the moments estimator for the cross-semivariogram applied to SOC and SON. This approach will help to identify observations that exert significant influence on the data, covering all available lags. The cartographic scheme representing the sampled locations is shown in Figure 7.

The dataset has been made possible thanks to the financial contributions of S.C. Johnson and Sons Inc., Madison Gas and Electric Corp. (MGE), and The Barker Fund, made possible through the UW-Madison Foundation and the College of Agricultural and Life Sciences.

Using the Bivariate hairplot with cross-semivariogram, shown in Figure 8, applied by the moments estimator to the data, we observed that samples #10, #85, #124, #88, and #6 stand out significantly, as indicated by the line red when considering the lag $h = 1$. When expanding the analysis to $h = 2$, observations #10 and #85 continue to be highlighted, but a new observation, #110, also shows influence. For $h = 3$, only observations #10 and #85 remain influential. Therefore, it is evident that observations #85 and #10 are consistently influential when considering the first three lags.

However, if we consider all lags, but fix a value of ζ, we can construct Andrews curves based on the perturbation process. Figure 9 shows all variations of $\zeta(-3, -2, -1, 1, 2, 3)$.

A thorough examination of the graphs reveals that observations #85 and #10 are consistently evident across all graphs, irrespective of the fixed value of zeta.

In considering the proposed matrix, $\mathbf{\Gamma}_\top$, it is possible to visualize a single graph of the Andrews curves, taking into account all lags and all ζ perturbations. Thus, Figure 10 represents the data matrix $\mathbf{\Gamma}_\top$:

We can again observe the prominence of the three observations in the graphs in which the ζ values were fixed. Thus, we can conclude that observations #85 and #10, are globally influential in the database under study, while observations #88, #6, #124 and #110 are locally influential, appearing as influential in just a few lags.

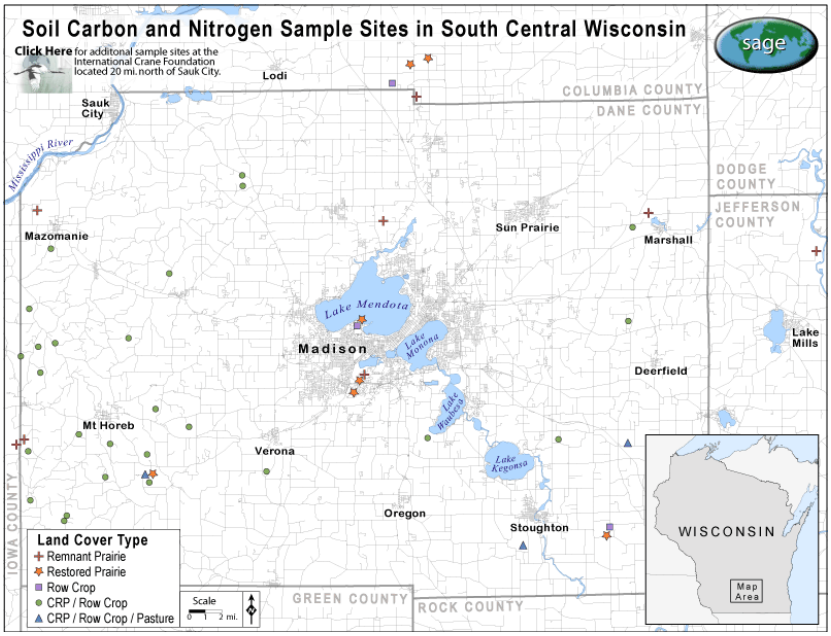
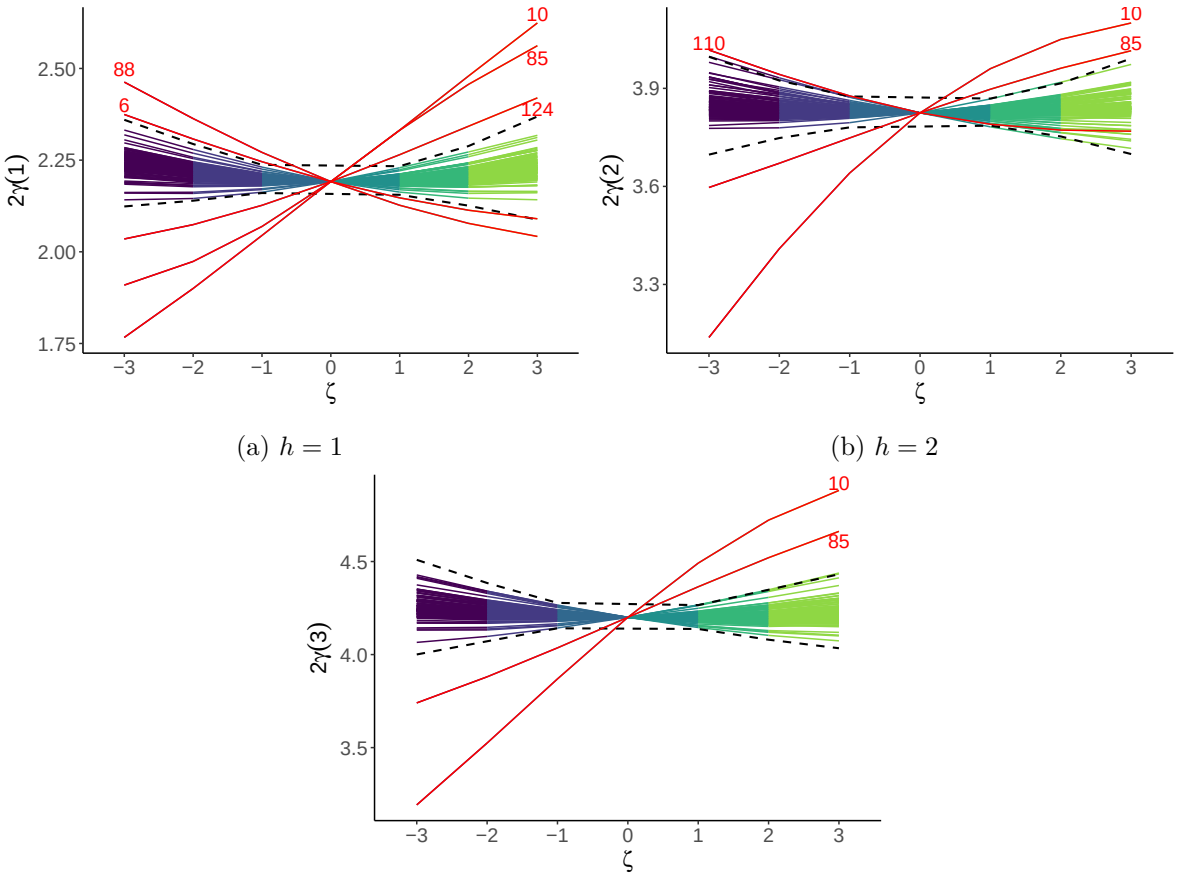


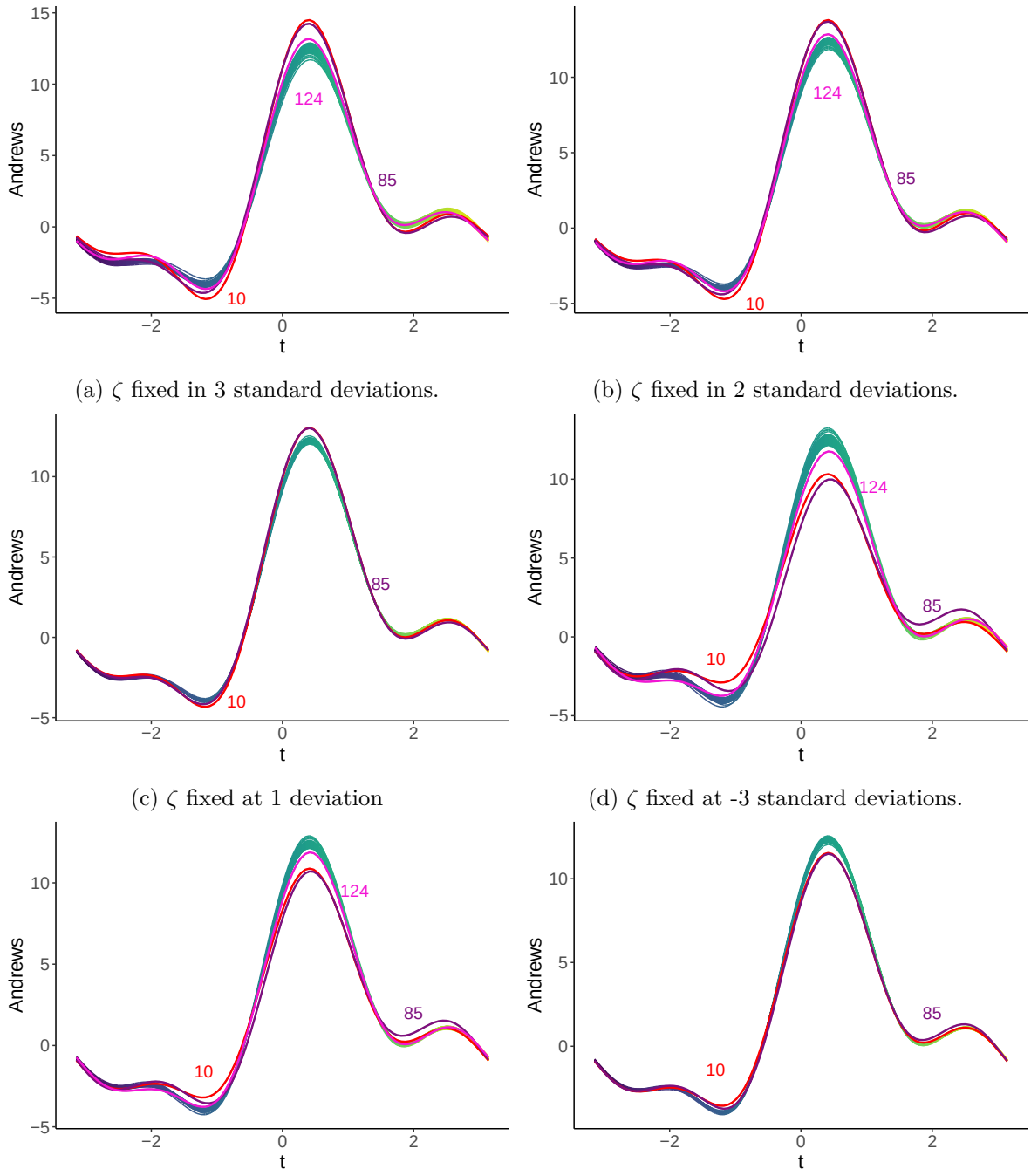
Figure 7: Soil carbon and nitrogen sample sites in south central Wisconsin. Source: https://sage.nelson.wisc.edu/soil-carbon-nitrogen/index.php



(a) $h = 1$ (b) $h = 2$

(c) $h = 3$

Figure 8: Bivariate hairplot considering only the first three lags for real data



(a) ζ fixed in 3 standard deviations.

(b) ζ fixed in 2 standard deviations.

(c) ζ fixed at 1 deviation

(d) ζ fixed at -3 standard deviations.

(e) ζ fixed at -2 standard deviations.

(f) ζ fixed at -1 standard deviations.

Figure 9: Hairplot based on Andrews curves for carbon and nitrogen data

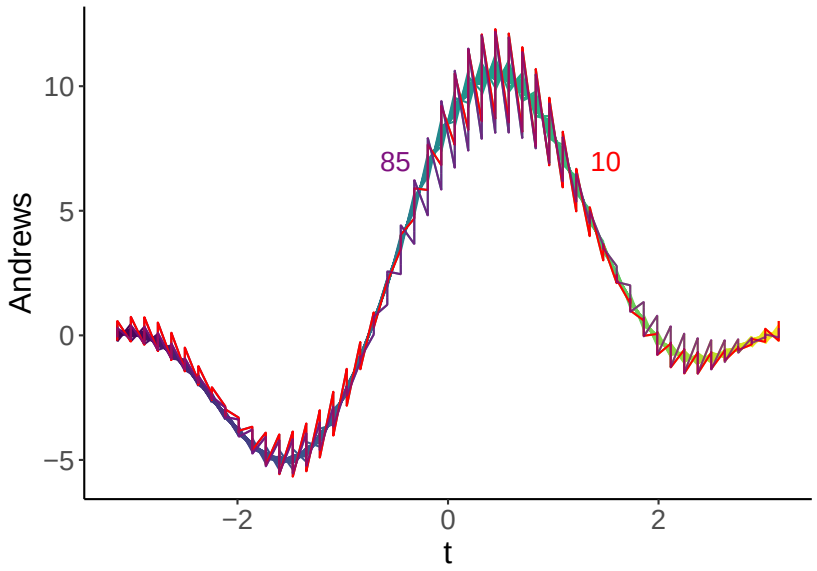


Figure 10: Hairplot based on Andrews curves for all ζ for carbon and nitrogen data

Figure 11 reveals two observations that are worthy of further examination: observation #10 and observation #85. Observation #10 is particularly notable in the Prairie Remnant agricultural area, while observation #85 stands out in the Forest agricultural area. Additionally, we observe other points that appear prominent; however, these are merely common outliers and do not exert significant influence on the results, thus not classified as influential observations.

To verify whether the potentially influential observations detected in the hairplot are indeed influential, we calculate boundary curves for the semivariogram at each lag using the bootstrap method Rao and Wu (1988). Next, we remove each of these observations identified as potentially influential to assess if their exclusion results in a significant change, such that the new semivariance, without the observation, falls outside the confidence interval constructed with all observations. Table 1 presents the removed observations, along with the semivariance estimates and corresponding boundary curves.

Table 1: Semivariance estimates and boundary curves after removal of influential observations

Obs	**Variogram**	**IC (LI)**	**IC (LS)**
Lag 1			
6	2.07793	2.115646	2.267315
10	1.493785	2.115646	2.267315
85	1.510226	2.115646	2.267315
88	2.038668	2.115646	2.267315
124	2.027754	2.115646	2.267315
Lag 2			
10	1.024996	3.730668	3.919488
85	3.311883	3.730668	3.919488
110	3.971344	3.730668	3.919488
Lag 3			
10	3.035415	4.104394	4.295898
85	2.13584	4.104394	4.295898

Table 1 confirms that all the observations identified in the hairplot are indeed influential. In particular, observations #10 and #85 stand out, as their removal results in drastic changes in the semivariance estimates.

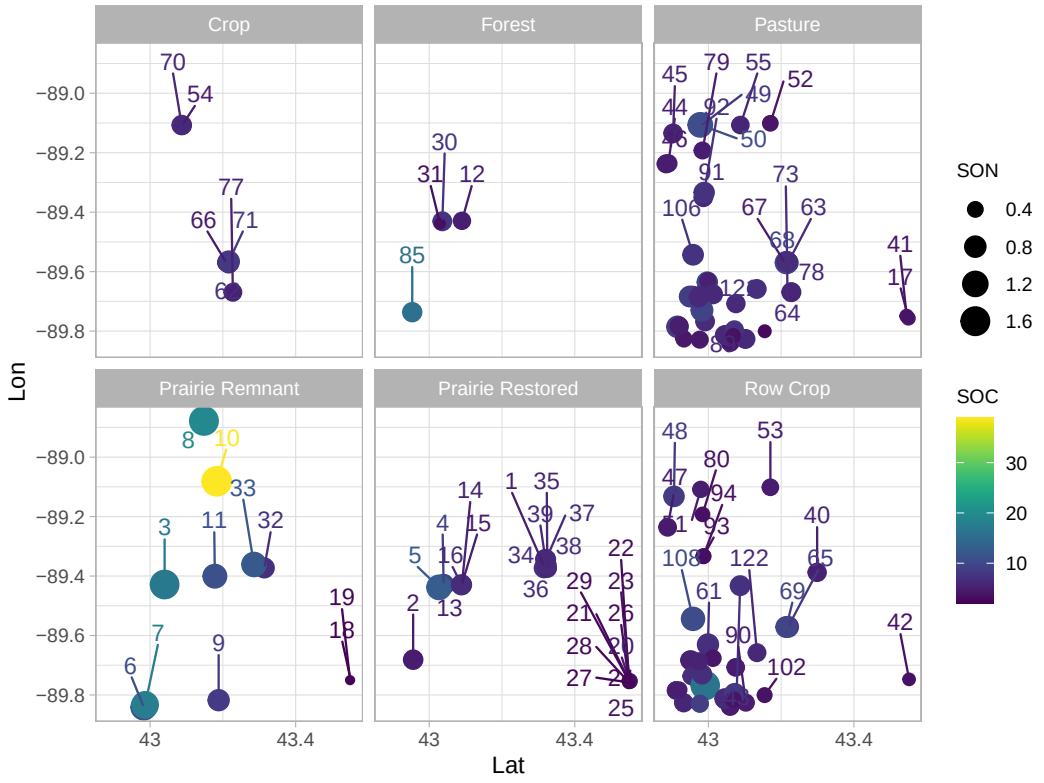


Figure 11: Distribution of observations across different land management areas

7. Conclusions

This study advances the analysis of spatial data in different contexts by integrating and extending established diagnostic techniques. By enhancing bivariate hairplots with boundary curves and combining them with cross semivariograms and Andrews curves, we provide a more automatic and interpretable framework for detecting influential observations in spatial datasets.

The introduction of statistically grounded boundary curves within the hairplot framework enables objective thresholds for identifying truly influential points, addressing limitations of subjective interpretation. The use of Andrews curves further complements this by allowing simultaneous visualization of all lags, offering richer insight into both local and global analysis, an essential distinction for understanding soil carbon and nitrogen dynamics.

The consistent identification of globally influential observations (e.g., #85 and #10) alongside locally influential ones (e.g., #88, #124, #6, and #110) demonstrates the method's capacity to disentangle different scales of influence. In particular, the stability of the results across perturbation scenarios underlines the reliability of the proposed approach.

Although alternative dimensionality reduction techniques like PCA could be employed, Andrews curves preserve individual observation signatures, aligning better with the diagnostic intent of the hairplot and enhancing interpretability.

Overall, this combination of techniques, coupled with reproducible `R` code, offers researchers and practitioners a transparent, replicable, and statistically rigorous tool for spatial influence detection. Such methodological improvements contribute valuable insights for sustainable land management strategies in southern Wisconsin and can be adapted to similar studies in diverse environmental settings.

All analyses and visualizations presented in this paper are fully reproducible. The `R` code is provided as Supplementary Material.

Acknowledgments

The authors gratefully acknowledge the financial support provided by the Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE) through the doctoral scholarship, grant IBPG-1723-1.02/21; by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) through the Sandwich Doctorate Program (PDSE), process 88881.981454/2024-01, which supported the co-supervision period at the Pontificia Universidad Católica de Chile; and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) through grants 306561/2020-4 and 404872/2023-9.

The authors also thank the Editor, the Associate Editor, and the anonymous referee for their valuable comments and suggestions, which substantially improved the quality of the manuscript.

References

Alcacer A, Epifanio I (2024). "Outlier Detection of Clustered Functional Data with Image and Signal Processing Applications by Archetype Analysis." *PLOS ONE*, **19**(11), 1–23. doi:10.1371/journal.pone.0311418.

Andrews DF (1972). "Plots of High-Dimensional Data." *Biometrics*, **28**(1), 125–136. doi:10.2307/2528964.

Babakhani M, Deutsch CV (2012). "Standardized Pairwise Relative Variogram as a Robust Estimator of Spatial Structure." *CCG Annual Report*, **14**.

Cressie N (2015). *Statistics for Spatial Data*. John Wiley & Sons. ISBN 978-1-119-11517-5.

Cressie N, Hawkins DM (1980). "Robust Estimation of the Variogram: I." *Journal of the International Association for Mathematical Geology*, **12**(2), 115–125. doi:10.1007/BF01035243.

Genton MG (1998). "Variogram Fitting by Generalized Least Squares Using an Explicit Formula for the Covariance Structure." *Mathematical Geology*, **30**(4), 323–345. doi:10.1023/A:1021733006262.

Genton MG, Ruiz-Gazen A (2010). "Visualizing Influential Observations in Dependent Data." *Journal of Computational and Graphical Statistics*, **19**(4), 808–825. doi:10.1198/jcgs.2010.09101.

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986). *Robust Statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York. ISBN 0-471-82921-8.

Jiménez-Varón CF, Harrou F, Sun Y (2024). "Pointwise data depth for univariate and multivariate functional outlier detection." *Environmetrics*, **35**(5), e2851.

Lark RM (2003). "Two Robust Estimators of the Cross-Variogram for Multivariate Geostatistical Analysis of Soil Properties." *European Journal of Soil Science*, **54**(1), 187–202. doi:!0.1046/j.1365-2389.2003.00506.x.

Matheron G (1963). "Principles of Geostatistics." *Economic Geology*, **58**(8), 1246–1266. doi:10.2113/gsecongeo.58.8.1246.

Moustafa RE (2011). "Andrews Curves." *Wiley Interdisciplinary Reviews: Computational Statistics*, **3**(4), 373–382. doi:10.1002/wics.160.

Ojo OT, Anta AF, Genton MG, Lillo RE (2023). "Multivariate Functional Outlier Detection Using the Fast Massive Unsupervised Outlier Detection Indices." *Stat*, **12**(1), e567. doi:10.1002/sta4.567.

Qu Z, Dai W, Euan C, Sun Y, Genton MG (2025). "Exploratory Functional Data Analysis." *TEST*, **34**, 459–482. doi:10.1007/s11749-024-00952-8.

Rao JNK, Wu CFJ (1988). "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association*, **83**(401), 231–241. doi:10.1080/01621459.1988.10478591.

Sun Y, Genton MG (2011). "Functional Boxplots." *Journal of Computational and Graphical Statistics*, **20**(2), 316–334. doi:10.1198/jcgs.2011.09224.

Yao Z, Dai W, Genton MG (2020). "Trajectory Functional Boxplots." *Stat*, **9**(1), e289. doi:10.1002/sta4.289.

Affiliation:

Fernanda De Bastiani
Statistics Department
Federal University of Pernambuco
50740-540, Av. Jorn. Aníbal Fernandes, Recife-PE, Brazil
E-mail: fernanda.bastiani@ufpe.br

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 55, Issue 3
doi:10.17713/ajs.v55.i3.2274
http://www.ajs.or.at/
http://www.osg.or.at/
Submitted: 2025-07-10
Accepted: 2026-04-21