Vigilada Mineducación

ASSESSING THE EFFECTS OF MULTIVARIATE FUNCTIONAL OUTLIER IDENTIFICATION AND SAMPLE ROBUSTIFICATION ON IDENTIFYING CRITICAL PM2.5 AIR POLLUTION EPISODES IN MEDELLÍN, COLOMBIA

LUIS MIGUEL ROLDÁN ALZATE

Tesis de grado

Asesor

Francisco Iván Zuluaga

UNIVERSIDAD EAFIT

ESCUELA DE CIENCIAS

MAESTRÍA EN MATEMÁTICAS APLICADAS

MEDELLÍN

Assessing the effects of Multivariate Functional outlier identification and sample robustification on identifying critical PM2.5 air pollution episodes in Medellín, Colombia.

Luis Miguel Roldán-Alzate[1*] and Francisco Zuluaga[1†]

[1*]Department of Mathematical Sciences, Universidad EAFIT, Carrera 49 N° 7 Sur-50, Medellín, 050022, Antioquia, Colombia.

*Corresponding author(s). E-mail(s): lroldan4@eafit.edu.co;
Contributing authors: fzuluag2@eafit.edu.co;
†These authors contributed equally to this work.

Abstract

Identification of critical episodes of environmental pollution, both as a outlier identification problem and as a classification problem, is a usual application of multivariate functional data analysis. This article addresses the effects of robustifying multivariate functional samples on the identification of critical pollution episodes in Medellín, Colombia. To do so, it compares 18 depth-based outlier identification methods and highlights the best options in terms of precision through simulation. It then applies the two methods with the best performance to robustify a real dataset of air pollution (PM2.5 concentration) in the Metropolitan Area of Medellín, Colombia and compares the effects of robustifying the samples on the accuracy of supervised classification through the multivariate functional DD-classifier. Our results show that 10 out of 20 methods revised perform better in at least one kind outliers. Nevertheless, no clear positive effects of robustification were identified with the real dataset.

Keywords: Air pollution, Multivariate Functional Outlier Detection, $\boldsymbol{\alpha}$-trimming, Sequential Transformations

1 Introduction

Functional data analysis is a field of study where data are functions rather than points on a finite-dimensional space. In the univariate case, points are functions defined on a continuous domain, from which we have discrete points that could be approximated through a smoothing procedure (Ramsay & Silverman, 2005). In the case of multivariate functional data, points are vectors of finite dimension with infinite-dimensional functions as elements (Berrendero, Justel, & Svarc, 2011).

Urban outdoor pollution is a very fertile field of study for the application of functional data analysis. Applications include spatio-temporal modeling of O_3 (Wang, Xu, & Li, 2020), outlier detection on industrial coal plants (Sánchez-Lasheras, Ordóñez-Galán, García-Nieto, & García-Gonzalo, 2020) and pollution-based spatial regionalization based on functional data methods (Liang, Zhang, Chang, & Huang, 2020). One of the most frequent applications of functional data analysis focused on air pollution is outlier detection (Febrero, Galeano, & González-Manteiga, 2008; Febrero-Bande, Galeano, & González-Manteiga, 2007; Martínez et al., 2014; Sánchez-Lasheras et al., 2020; Shaadan, Deni, & Jemain, 2012; Shaadan, Jemain, Latif, & Deni, 2015; Torres, Nieto, Alejano, & Reyes, 2011; Torres et al., 2020). Even when there are many applications of outlier identification in functional data analysis for air pollution data, there is an opportunity for analyzing multivariate functions. Among the articles identified, Martínez et al. (2014); Sánchez-Lasheras et al. (2020); Shaadan et al. (2015) are devoted to more than one variable, but they address each variable individually.

Outdoor concentrations of thin particulate matter (PM2.5) are particularly relevant from the perspective of public health, since exposure to high concentrations of this pollutant is proven to be related to a higher risk of death (World Health Organization, 2006). The Metropolitan Area of Medellín, in Colombia, has gone through peaks in PM2.5 daily average concentrations, and there have been efforts both from the city government and the metropolitan environmental authorities to cope with this issue.

Taking advantage of the presence of many PM2.5 measuring stations, this article applies multivariate functional depth-based techniques to average daily concentrations of PM2.5. Depth estimation, classification and outlier detection methods for multivariate functional data are considered.

There are plenty of methods for outlier detection and a lack of comprehensive reviews, the most comprehensive one being Ieva and Paganoni (2017) , which leads to an opportunity to address this issue and complement it with the combination of methods. Through simulation, this article tests available methodologies and complements the review with combinations of them, mainly an application of univariate α-trimming methods to multivariate functional samples and an application of multivariate functional boxplots with different depth measures.

The article has the following sections: Section 2 - Methods, describes multivariate functional depth measures, outlier detection and the simulation

procedures. Section 3 presents the results of the analysis applied to the PM2.5 dataset. Section 4 presents discussion and concluding remarks.

2 Methods

According to Febrero et al. (2008), outliers in functional data are either the result of a data gathering mistake or the consequence of a disturbed data generating process. As a consequence, they may induce bias, on one hand, or be a potential source of knowledge about the disturbances that provoked their appearance, on the other (Febrero et al., 2008).

Methods for outlier identification are based on the notion of statistical depth, that relates to identifying how central is a point $x \in \mathbb{R}^d$ with respect to a probability distribution F (Liu, Parelius, & Singh, 1999). A depth function can be defined as any function $D(x; P)$ that provides an ordering from the center outwards to the points $x \in \mathbb{R}^d$ based on the probability measure P (Zuo & Serfling, 2000). Measures of depth result from applying depth functions to examples from a probability distribution.

In the univariate functional scenario, the space in which both each individual point and the whole sample are considered is the space of continuous functions in a specific domain (Ramsay & Silverman, 2005). There are many univariate functional depth functions and measures, such as Fraiman-Muniz depth (Fraiman & Muniz, 2001), h-modal depth (Cuevas, Febrero, & Fraiman, 2006), random projection depth (Cuevas, Febrero, & Fraiman, 2007), band and modified band depth (López-Pintado & Romo, 2009), simplicial band depth (López-Pintado, Sun, Lin, & Genton, 2014), halfspace depth and random Tukey depth (Cuesta-Albertos & Nieto-Reyes, 2008). Each approximation highlights different attributes of the curves, which leads to different ordering and differences in final results. Multivariate functional depth can be built as a combination of multivariate depths over the continuous domain of the data, or as a combination of functional depths over the multivariate scenario.

Multivariate functional depth measures are either combinations of univariate functional depth measures to fit a multivariate shape (as in Ieva and Paganoni (2013)) or combinations of multivariate measures to fit a functional shape (as in Claeskens, Hubert, Slaets, and Vakili (2014)).

Figure 1 shows the analytical structure this section. Firstly, it describes aggregation methods. Secondly, it summarizes the depth measures that are most commonly aggregated by each method. Afterwards, it covers outlier identification methods. As said in the introduction, mentions of supervised classification are left to the last section of this document, regarding the application of the methodology to air pollution.

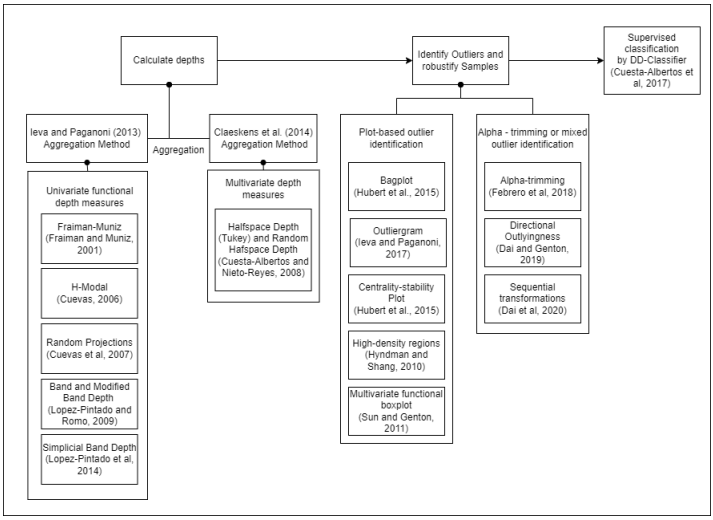


Fig. 1: Process of robustification tested on this article.

2.1 Notation

Functional data are defined henceforth as continuous functions Y_i defined on a compact interval U from which we observe T units with $t_1, t_2, ..., t_T \in U$, so that $Y_i : U \to \mathbb{R} : t \to Y_i t$. If we consider K variables, with $k = 1, 2, 3, ..., K$, there will be a K-variate functional vector and $Y_i = (y_{i_1}, y_{i_2}, ..., y_{i_K})$ univariate functional data points over a continuous domain delimited by the interval U. Each multivariate functional observation Y_i can also be represented as $Y_i = (y_{i_1}, ..., y_{i_T})$ K-variate points. In this context, each y_{i_t} can be viewed a realization of a random vector, and each y_{i_k} can be viewed as a realization of a univariate functional variable. If each multivariate functional data point Y_i is viewed as a realization of a multivariate functional process, the complete set of multivariate functions is named Y and comprehends N multivariate functional observations; respectively, the cross-sectional view of Y for a specific time point t is Y_t and the univariate functional view of Y for a specific variable k is Y_k.

There are two possibilities for calculating the multivariate functional depth with respect to a set Y, $D(Y_i; Y)$: to consider the multivariate case at first, and after that the functional case or, in the contrary, to consider first the univariate functional case and then the multivariate case. Given a multivariate depth function $Dm(y_{i_t}; Y_t)$, the first case can be defined as:

$$D(Y) = \mathrm{FA}_U(\mathrm{Dm}(y_{i_t}; Y_t)) \tag{1}$$

Where FA is a functional aggregation of Dm, which is a multivariate depth, over the domain U. In the second case, given an univariate functional depth function $Df(y_{i_k}; Y_k)$, the multivariate functional depth function can be defined as:

$$\mathrm{D}(Y) = \mathrm{MA}_K(\mathrm{Df}(y_{i_k}; Y_k)) \tag{2}$$

Where MA is a multivariate aggregation of Df, which is an univariate functional depth, over the variables K.These two definitions are building blocks for the more specific definition of multivariate functional depth.

2.2 Aggregation methods to build multivariate functional depths

As depicted in Figure 1, two aggregation methods were found in the review. The first method, stated by Claeskens et al. (2014) aggregates multivariate depths whilst the second, stated by Ieva and Paganoni (2013) aggregates functional depths.

2.2.1 Method 1: from multivariate to multivariate functional depth:

The method proposed by Claeskens et al. (2014) "averages a multivariate depth function over time points, but in addition it includes a weight function" (P.412). The multivariate functional depth proposed by the authors can be defined, starting from (1), as follows [1]:

$$\mathrm{MFD}(y_{i_t}; Y_t) = \sum_{t=1}^{T} \mathrm{Dm}(y_{i_t}; Y_t) * \mathrm{w}(t)$$

where $w(t)$ is a function defined on the interval U which values sum up to 1 over its domain and weights the magnitude of the multivariate distribution at each point t. Weighting functions can depend both on t and on the characteristics of F_{Y_t} [2]. The original design from Claeskens et al. (2014) is based on random projections multivariate depth functions. Nevertheless, the assembling process can be executed with any multivariate depth function, and many of them will be discussed in the next section.

Algorithm 1 Pseudo code of Method 1

for $t \in U$ **do**
  Calculate $\mathrm{Dm}(Y_{i_t}; F_{Y_T})$
  Calculate $\mathrm{w}(t)$
end for
$\mathrm{MFD}(y_{i_t}; Y_t) = \sum_{t=1}^{T} \mathrm{Dm}(y_{i_t}; Y_t) * \mathrm{w}(t)$

[1] The notation is standardized aiming uniformity and it is not necessarily equal to the original
[2] The original weighting function proposed by Claeskens et al. (2014) gives to each point t a weight that corresponds to the proportion of amplitude of the multivariate dataset at that point.

2.2.2 Method 2: from functional depth to multivariate functional depth:

Method 2, proposed by Ieva and Paganoni (2013) consists in calculating univariate functional depths and then aggregating them by weighted averages to give an estimate of multivariate functional depth. Starting from an univariate functional depth function $\mathrm{Df}(y_{i_k}; Y_k)$ that assigns a depth for every y_{i_k} function with respect to a set Y_k, the method can be stated, starting from (2) as

$$\mathrm{MFD}(Y_i) = \sum_{k=1}^{K} \mathrm{w}(k)\mathrm{Df}(y_{i_k}; Y_k)$$

Where $\mathrm{w}(k)$ is a nonnegative weighting function or parameter which values summate to one and that assigns a weight to each dimension k under consideration.

Algorithm 2 Pseudo code of Method 2

for $k \in K$ **do**
  Calculate $\mathrm{Df}(Y_{i_k}; F_{Y_K})$
  Calculate $\mathrm{w}(k)$
end for
$\mathrm{MFD}(Y_i) = \sum_{k=1}^{K} \mathrm{w}(k)\mathrm{Df}(y_{i_k}; Y_k)$

Being depth functions that were created specifically for functional data, band depth and modified band depth are the main depth functions that fit this method according to Ieva and Paganoni (2013), but other univariate functional depth functions can also be used with this aggregation framework. The following section illustrates some of the functional depths that can be applied.

2.3 Depth measures

This section summarizes depth measures that are building blocks to both methods listed above. They are divided in two groups: multivariate depth measures used for method one, and univariate functional depth measures, commonly used for method two.

2.3.1 Multivariate depth measures: Multivariate Halfspace Depth[3]

Halfspace depth or Tukey's depth is a measure of centrality for multivariate data. Following our notation, for a random variable $y_{i_t} \in \mathbb{R}^K$ belonging to a sample Y_t and any vector $u \in \mathbb{R}^{\mathbb{K}}$, halfspace depth can be defined as (Claeskens et al., 2014):

[3]Simplicial band depth is also used with Method one, as can be seen in López-Pintado et al. (2014), but that method is not explored in this document

$$\text{HD}(y_{i_t}; Y_t) = \frac{1}{N} \min_{u \in \mathbb{R}^K, \|u\|=1} \#\{Y_{t_n}, n = 1, ..., N : u' Y_{t_n} \geq u' y_{i_t}\}$$

i.e. the minimum proportion of multivariate points of the cross-sectional set $Y_t = y_{t_1}, ..., y_{t_n}$ covered by a halfspace -a projection through u.

2.3.2 Univariate functional depth measures

Band and modified band depth

For a set of functions $y_1, y_2, ..., y_K$, a graph can be defined as $G(y_k) = \{(t, y_{k_t}) : t \in U\}$. Based on this detinition, López-Pintado and Romo (2009) define a band as

$$\text{B}(y_{1_k}, y_{2_k}, ..., y_{N_k}) = \{(t, y_k) : t \in U, y = \alpha * \min_{r=1,...,n} y_{kt_r}(t) + (1-\alpha) * \max_{r=1,...,n} y_{kt_r}(t)\}$$

With $\alpha \in [0, 1]$. It is, the area contained between the maximum and the minimum of a set of graphs for each t. From this definition, band depth is defined as the proportion of bands that fully contain a specific curve y_{i_k}

$$\text{BD}_N^{(j)}(y_{i_k}) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < ... < i_j \leq n} I\{X(t) \subseteq B(y_{i_k 1}, y_{i_k 2}, ..., y_{i_k j})\}$$

Starting from this definition, modified band depth can be defined as the weighted average of the centrality of each discretization point. This function measures univariate depth at each discretization point and aggregates the measures to get univariate functional measures.

Fraiman-Muniz depth

Taking advantage of the definition of simplicial depth (Liu, 1990), Fraiman and Muniz Fraiman and Muniz (2001) build an univariate functional depth defined as

$$\text{FMD}(y_{i_k}) = \int_U \left| \frac{1}{2} - F_t(y_{i_k t}) \right|$$

where $F_t(y_{i_k t})$ is the empirical distribution of the univariate sample gathered from y_{kt} evaluated at point i

H-Modal depth

Taking advantage of Kernel density estimation, and identifying the mode of a sample as the most densely surrounded observation, Cuevas et al. (2006)

develop a notion of depth to approach the concept of mode for functional data. H-Modal depth function can be defined as

$$\text{HMD}(y_{i_k}; h) = g(y_{i_k}; h; y_{1_k}, ..., y_{n_k}) = \sum_{j=1}^{n} K_h(\|y_{i_k} - y_{j_k}\|)$$

Random projections depth

Projection depth can be defined as the univariate depth of a unidimensional projection of a point y_{i_k} belonging to a multivariate or functional process Y_k. The projection is defined as the inner product $\langle a, X \rangle = \int_0^1 a(t)X(t)dt$ (Cuevas et al., 2007). Since a single projection could be highly biased, Cuevas et al. (2007) use 50 random directions to project the data and average the depths estimated on each projection to get a global depth.

2.4 Outlier detection

Outlyingness can be seen as the inverse of depth, it is, an ordering from the surface inwards. Many methods have been proposed for depth-based outlier identification (Ieva & Paganoni, 2017). We describe them as general frameworks in the following paragraphs.

Outlier detection based on bagplots

Hubert, Rousseeuw, and Segaert (2015) propose a method for outlier identification in multivariate functional data that combines the approach to multivariate functional depth provided by Claeskens et al. (2014) and the notion of bagplot for multivariate data developed by Rousseeuw and Ruts (1999).

The starting point of bagplots is the definition of an α- depth region as the region that contains points with at least an α level of depth, i.e. $D_\alpha = \{x \in \mathbb{R}^p; D(Y_i; P_Y) \geq \alpha\}$ with α-depth contour being the boundary of the set D_α. This concept allows for the definition of a halfspace median as the center of gravity of the smallest halfspace depth region[4]. The bag of a bagplot is defined as the convex hull that contains 50% depth region. The outlier criterion is based on inflating the bag by a factor of three and identifying points that rely outside of the bag as outliers.

In this article, we build a K-variate halfspace-based bagplot at each time point t to identify multivariate outliers at each discrete point using the bagplot criterion. An observation is identified as an outlier if it outlies at least at one discretization point.

Outlier detection based on adjusted outlyingness -Centrality-stability plot-

The Centrality-stability plot is the an outlyingness detection method proposed by Hubert et al. (2015), based on the multivariate functional skew-adjusted

[4]Even when Hubert et al. (2015)'s definition of depth region is based on the halfspace depth, any other depth function that meets the properties enumerated above can be used to build depth regions. This is stated but not demonstrated on Hubert et al. (2015)

projection depth (MFSPD), with Claeskens et al. (2014) aggregation method as building block. There is a starting point definition of the skew-adjusted projection depth based on the notion of adjusted outlyingness. The authors define Adjusted Outlyingness as

$$\mathrm{AO}(y_{i_t}; Y_t) = \sup_{\|v\|=1} \begin{cases} \frac{v' y_{i_t} - \mathrm{median}(v' Y_t)}{w_2(v'Y) - \mathrm{median}(v' Y_t)} & \text{if } v'x > \mathrm{median}(v' Y_t) \\ \frac{\mathrm{median}(v' Y_t) - v' u_{i_t}}{\mathrm{median}(v' Y_t) - w_1(v' Y_t)} & \text{if } v'x < \mathrm{median}(v' Y_t) \end{cases}$$

with w_1, w_2 defined as lower and upper parameter boundaries for the skew-adjusted outlyingness in Hubert et al. (2015).

Based on that definition, the authors define a *Multivariate Functional Skew-adjusted projection depth (MFSPD)*

$$\mathrm{MFSPD} = \sum_{j=1}^{T} \frac{1}{(1 + AO(X(t_j); P_n(t_j))W_j^{-1}}$$

The centrality-stability plot that gives the criteria for identifying outliers in Multivariate Functional Data is a scatter plot of the pairs given by

$$\left(1 - \mathrm{MFSPD}_n(Y_i; P_n); \underset{j}{ave}\left((1 + \mathrm{AO}(Y_i(t_j); P_n(t_j))W_j^{-1}\right) - \frac{T}{\mathrm{MFSPD}_n(Y_i; P_n)}\right)$$

For points that are isolated outliers, the centrality-stability plot will show relatively higher values on the y-axis. On the other hand, for persistent outliers, there would be a relatively low value for the y axis and a relatively high value of the X-axis.

Outlier detection based on multivariate functional outliergrams

The multivariate functional outliergram is a method based on the notion of multivariate functional (modified) band depth and the (modified) epigraph index for multivariate functional data (Ieva & Paganoni, 2017). Starting from the definition of epigraph index (EI) as the proportion of the curves that are above a specific function under consideration, and the modified epigraph index as weighted count of the proportion of the curves above a specific curve, the multivariate outliergram method takes advantage of the inequality

$$\mathrm{MBD}^J_{y_1,\ldots,y_n}(y_i) \leq a_o + a_1 MEI_{y_1,\ldots,y_n}(y_i) + a_2 n^2 (MEI_{y_1,\ldots,y_n}(y_i))^2$$

where $\mathrm{a}_0 = a_2 = -2/(n(n-1))$ and $a_1 = 2(n+1)/(n-1)$ [5]

[5]This inequality is shown in Ieva and Paganoni (2017) in more detail, with the corresponding plots of (modified) band depth (X axis) against (modified) epigraph index (Y axis). Every point falls below the aforementioned inequality, but points that fall too far from the boundary could be understood as shape outliers, while points whit low (M)BD could be considered magnitude outliers.

Specifically, Ieva and Paganoni (2017) compute a distance $d_i = a_0 + a_1 MEI_i + a_2 n^2 (MEI_i)^2 - MBD_i, i = 1, ..., n$ and define as outliers the data points with $d_i \geq Q_{d3} + 1.5 h_d IQR_d$, with Q_{d3} and IQR_d the third quartile and interquartile range of $d_1, ..., d_n$ and h_d is a function of a robust measure of data skewness, specifically for this case the exponential model developed upon the medcouple by Hubert and Vandervieren (2008).

Outlier detection based on α-trimming

The method of trimming for outlier identification consists in iteratively calculating functional depths, identifying a depth cutoff C and dropping observations with depths less than or equal to that cutoff point C. According to Febrero et al. (2008), there are two ways for defining the cutoff point C using bootstrap samples.

The first method obtains a trimmed sample by eliminating the most suspicious curves from the original sample. After that, depths must be calculated, B bootstrap samples of size n must be gathered and a threshold C^b as the empirical 1% percentile of the depths should be identified for each sample. The cutoff point for the original distribution would then be the median of the C^b values.

The second procedure omits the initial trimming and weights the bootstrapping procedure so that deeper points have a higher probability of being sampled. This second procedure is the one used in this article on $\alpha-$trimming. The sequence is as follows:

Algorithm 3 Pseudo code for α - trimming method

for $i \in 1 : n$ **do**
  Calculate multivariate functional depths $MFD(y_i)$
end for
for $i \in 1 : n$ **do**
  Calculate weights $w(y_i) = \frac{\text{depth}_{y_i} - \min(depth)}{\max(depth) - \min(depth)}$
end for

- Generate B weighted bootstrap samples with the vector w of weights determining the probability of a point of being sampled.
- Identify the depth value for the 99th percentile (c) for all samples.
- Find the median of the depth values and identify as outliers all observations with depths below this threshold.

Multivariate Functional Boxplot

Multivariate functional boxplot, developed by Ieva and Paganoni (2013) is an extension of the univariate functional boxplot proposed by Sun and Genton (2011). Similar to the univariate boxplot, the procedure consists on calculating the (modified band) depth for each multivariate functional point, then ranking

the functions according to their depths and defining the envelope that includes the α% (usually 50%) central region, inflating the central envelope by a factor of h (usually 1.5), and then marking as outliers those functions that are out of the inflated region at least at one discretization point.

Functional High Density Region boxplot

HDR for functional data is a bivariate HDR boxplot (Hyndman & Shang, 2010) that plots the depths of the first two principal components scores for each observation.

The first step for this method is building a bivariate dataset conformed by the depths of the two first principal components for each observation. After that, a kernel density is to be estimated for those depths, according to the following equation:

$$\hat{\mathrm{f}}(z) = \frac{1}{n}\sum_{i=1}^{n} K_{h_i}(z - Z_i)$$

where Z_i is a set of bivariate points for the ith dimension, K is the kernel function and h_i is the bandwidth.

The High Density Region - HDR is defined as

$$\mathrm{R}_\alpha = \{z : \hat{f}(z) \geq f_\alpha\}$$

where f_α is the region with probability coverage $1 - \alpha$. The points displayed in the HDR boxplot as the mode, defined as $\text{argsup}\hat{f}(z)$, and also the 50% and 99% higher density regions. The points outside HDR are considered outliers.

Directional outlyingness

Dai and Genton (2019) develop the notion of directional outlyingness, defined both as a method for integrating point-wise multivariate functional as a measure of outlyingness that goes beyond depth and tries to capture both magnitude and direction of outlyingness. Multivariate directional outlyingness is defined by Dai and Genton (2019) as

$$\mathbf{O}(y_{i_t}) = \mathbf{O}(y_{i_t}) * \mathbf{v}(t)$$
$$\mathbf{O}(y_{i_t}) = \{1/d(y_{i_t}) - 1\} * \mathbf{v}(t)$$

where $d(y_{i_t}) > 0$ is a depth measure, and $\mathbf{v}(t) = \{y_{i_t} - \mathbf{Z}(t)/\|y_{i_t} - \mathbf{Z}\|\}$, where $\mathbf{Z}(t)$ is the unique median of the empirical distribution, as measured by depth $d(y_{i_t})$. The multivariate functional version of this measure of outlyingness is generated based on Method 1.

The outlier identification procedure is as follows(Dai & Genton, 2019):

1. Generate a bivariate set of both the directional outlyingness for an observation and its variation from the mean outlyingness. According to Dai and Genton (2019), the distribution of this is approximatedly bivariate normal.
2. Calculate the robust Mahalanobis distance for this dataset using de Minimum Covariance Determinant Method.

3. Approximate the tail of the distribution.
4. Identify as outliers curves that are grater than a cutoff point C

Method of sequential transformations

Dai, Mrkvička, Sun, and Genton (2020) offer a different perspective to account for both magnitude and shape outlyingness. The algorithm consists on iteratively identifying magnitude outliers through an effective magnitude-based method such as α-trimming or functional boxplot and then applying transformations to the data in order to get shape outliers. There are no standard criteria of how many transformations -or iterations- are to be considered or which depth function is preferable. The algorithm proposed by Dai et al. (2020) is as follows:

Algorithm 4 Pseudo code for secuential transformations method

- Identify *magnitude* outliers, it is, outliers that are separated from the sample across the complete domain U.
- Transform data to get a modified sample.
- Repeat step 1. Identified outliers are shape outliers.
- Apply a second transformation
- Repeat step 1 Identified outliers are shape outliers.

According to Dai et al. (2020), two sequences of transformations that can be used in the univariate functional case, but extensible to the multivariate functional case, are:

Transformation sequence 1 - ('O' Transformation)

1. Creating a curve of multivariate outlyingness at each point in time.
2. Creating a curve of multivariate outlyingness at each point in time to the previous curve.

In this sequence, a curve of outlyingness is the functional representation of the multivariate outlyingness -defined as the inverse of depth- of a multivariate point at each point in time t. In this case, we use directional outlyingness, as implemented in Ojo, Lillo, and Fernandez Anta (2020).

Transformation sequence 2 - ('T' Transformation)

1. Taking first order derivatives
2. Taking second-order derivatives

2.5 Simulation and testing of summarized approaches

In this section, we simulate five kinds of outlying bivariate functional data samples from the same process and, following the methodology implemented by Ieva and Paganoni (2017), we simulate 100 data sets with the same structure

and evaluate their average true positive and false positive outlier identification rate, in order for us to identify the mechanism that most precisely identify the known outliers. On doing so, the following innovations are made to complement Ieva and Paganoni (2017)'s work, specifically on data simulation and the number of algorithms tested:

- We add a two different types of outlyingness, obeying Hubert et al. (2015) distinction between persistent and isolated outliers, and Nagy, Gijbels, and Hlubinka (2017) suggestion of trend outliers.
- We complement boxplot outlier identification (originally made with (Modified) Band Depths) with H-modal, random projections and Fraiman-Muniz, in order for us to compare different depth structures.
- We added α trimming identification with Band Depth, Modified Band Depth, hmodal, Fraiman-Muniz and Random Projections depths.
- We added directional outlyingness and sequential transformations, which are not present on Ieva and Paganoni (2017).
- In the original article, Ieva and Paganoni (2017) adjust for low depths when using the Multivariate Functional Outliergram and show results only for the adjusted version. Here, for illustration, we show both methods.
- For the same purpose, we use Febrero et al. (2008) outlier identification mechanism with Modified Band Depth, H-modal, Random Tukey and Simplicial depth and compare the results. We do the same with sequential transformations.

We address six different types of outlyingness: magnitude, shape, covariance structure, mixed, trend and isolated mixed outliers. Each type of outlyingness is illustrated through a simulated dataset. The simulated dataset consists of two variables that come from a reference bivariate gaussian process. We also test for the behavior of the outlier identificator under a heavy-tail distribution. To this purpose, we use the same mean functions and simulate a t-distributed process with 3 degrees of freedom. We ran this simulations taking advantage of the function generate_gauss_mfdata from the R Package Roahd (Ieva, Paganoni, Romo, & Tarabelloni, 2019).

The generating process is the following:

$$(X, Y), X(t) = \mu_X(t) + \mathbb{Z}_X(t), Y(t) = \mu_Y(t) + \mathbb{Z}_Y(t)$$

where

$$\mu_X(t) = \sin(2\pi t), t \in I = [0, 1]$$
$$\mu_Y(t) = \sin(4\pi t), t \in I = [0, 1]$$

As in Ieva and Paganoni (2017), magnitude outliers consider an upwards shift on the parameter $\mu_X(T)$ in two units. Shape outliers consider an horizontal displacement of the curves so that the shape is modified. Covariance structure outliers consider a more highly correlated dataset for the outliers. Mixed combine shape and magnitude outliers.

For isolated mixed outliers, we modify the original structure so that points in the highest 10% of the domain follow the distribution of mixed outliers.

The details of the generating processes of the outliers can be found in (Ieva & Paganoni, 2017). For trend outliers we replace the mean function for an entirely different function over the same domain, following a quadratic structure on both cases.

Figure 2 shows the results of the simulation for a 20% of outlying observations.

We simulated 200 multivariate functional observations in 200 discretization points over the (0,1) domain, with 20% and 5% degrees of outliers contaminating the data. Following Ieva and Paganoni (2017), no smoothing procedure was implemented to the data. Also, to complement the approach made in Ieva and Paganoni (2017) and in order to test if the algorithms would perform equally well under heavy-tailed distributions, we estimated two different random processes for each simulation: a gaussian distribution and a t distribution with 3 degrees of freedom.

After the simulations, we implemented the outlier identification procedures enunciated before, and measured two indicators of accuracy: the true positive rate, which was the same used by Ieva and Paganoni (2017), and an indicator of accuracy that consists on comparing the true positive (TP) and false positive rate (FP) $TP+1/FP+1$, to adjust for the misclassification of false positives[6]. Table 2 shows complete results for true positives, and table 3 shows results for the true positive - false positive indicator.

Simulation results and discussion

Simulations show that there is no standard method for getting high levels of accuracy on all outlyingness types. Table 1 shows best peforming methods for each one of the tested distributions.

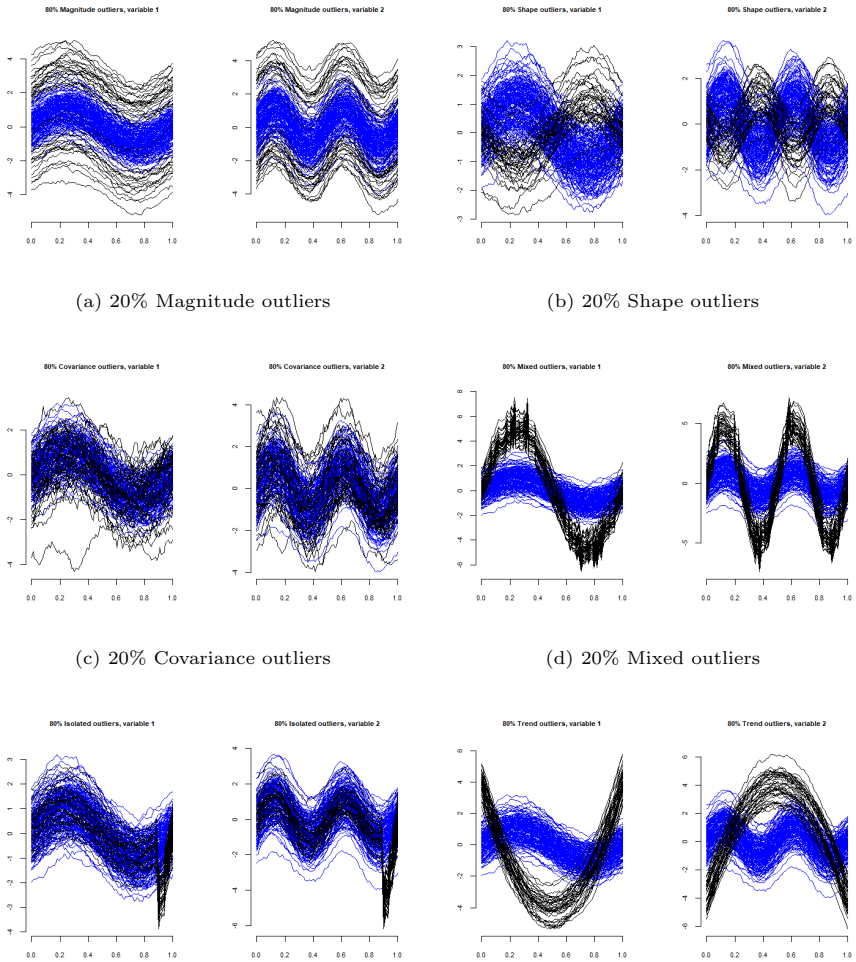
Outlier type	Normal 5%	Normal 20%	T distribution 20%
Magnitude	Centrality-Stability	Alpha-Hmodal	Alpha-Hmodal
Shape	Directional	Seq-O	Directional
Covariance	Seq-O	Alpha-BD	Directional
Mixed	Boxplot, Seq-T	Boxplot, Seq-T	Seq-T hmodal
Trend	Boxplot, Seq-T	Seq-T	Directional
Isolated	Seq-T	Seq-T	Seq-T hmodal

Table 1: Best performing method for each outlyingness type, according to TP-FP criterium

According to this results, there are differences on outlier identification not only with respect to the type of multivariate functional outlier, but also to the percentage of outlying observations and the dispersion of the generating process. Sequential transformations using T transformation twice performs well

[6]The indicator goes from 0.5, indicating the poorest performance where all positives are false and none of them are true, and 2, where all positives are true and none of them are false. This indicator, nevertheless, must be complemented with the false positive rate.

for Mixed and Isolated outliers for all generating processes, and for Trend outliers for normally distributed simulations, meaning that it performs better with outliers that are significatively anomalous in shape. Directional outlyingness performs better on very disperse distributions for trend, shape and covariance outliers, and for the less contaminated normal distribution for shape outliers.



(a) 20% Magnitude outliers

(b) 20% Shape outliers

(c) 20% Covariance outliers

(d) 20% Mixed outliers

(e) 20% Isolated outliers

(f) 20% Trend outliers

Fig. 2: Simulated data (blue) with outlying observations (black)

Another relevant result that can be seen on tables 2, 3, and 4, is that only 10 out of 20 methods display the best performance at least in one of

the cases studied. Boxplot, centrality-stability plot, outliergram, bagplots and high density regions did not have the best performance in any of the cases.

	Magnitude	Shape	Covariance	Mixed	Trend	Isolated
		20% Outliers				
boxplot-MBD	0.16	0.02	0.02	1.00	0.96	0.00
boxplot-hmodal	0.16	0.02	0.02	1.00	0.99	0.01
boxplot-FM	0.16	0.02	0.02	1.00	0.98	0.00
boxplot-RP	0.10	0.00	0.01	0.98	0.90	0.00
centrality.stability	0.73	0.01	0.22	0.00	0.20	0.02
outliergram	0.55	0.00	0.04	0.00	0.00	0.01
outliergram_m	0.60	0.40	0.36	0.97	0.96	0.05
bagplot	0.32	1.00	0.98	1.00	1.00	1.00
hdr	0.16	0.22	0.75	0.22	0.25	0.98
directional	0.61	1.00	0.92	1.00	1.00	0.17
seq-O	0.88	1.00	0.79	1.00	1.00	0.07
seq-T-MBD	0.16	0.25	0.21	1.00	1.00	0.40
seq-T-hmodal	0.16	0.36	0.24	1.00	1.00	1.00
seq-T-FM	0.16	0.22	0.15	1.00	1.00	1.00
seq-T-RP	0.11	0.18	0.19	1.00	1.00	1.00
alpha-MBD	0.75	0.09	0.18	0.40	0.41	0.11
alpha-BD	0.58	1.00	0.91	1.00	1.00	0.58
alpha-hmodal	0.97	0.63	0.55	0.68	0.70	0.24
alpha-FM	0.95	0.26	0.37	0.68	0.55	0.15
alpha-RP	0.94	0.26	0.37	0.67	0.56	0.15
		5% Outliers				
boxplot-MBD	0.34	0.07	0.05	1.00	1.00	0.00
boxplot-hmodal	0.33	0.07	0.05	1.00	1.00	0.23
boxplot-FM	0.30	0.04	0.03	1.00	1.00	0.00
boxplot-RP	0.27	0.03	0.03	1.00	0.99	0.00
centrality.stability	0.98	0.73	0.38	1.00	1.00	0.02
outliergram	0.59	0.67	0.21	0.97	0.98	0.01
outliergram_m	0.60	0.98	0.57	1.00	1.00	0.10
bagplot	0.22	1.00	0.98	1.00	1.00	1.00
hdr	0.06	0.00	0.72	0.00	0.00	0.98
directional	0.64	1.00	0.95	1.00	1.00	0.72
seq-O	0.99	1.00	0.88	1.00	1.00	0.21
seq-T-MBD	0.34	0.39	0.28	1.00	1.00	1.00
seq-T-hmodal	0.33	0.46	0.32	1.00	1.00	1.00
seq-T-FM	0.30	0.33	0.18	1.00	1.00	1.00
seq-T-RP	0.25	0.26	0.26	1.00	1.00	1.00
alpha-MBD	0.97	0.33	0.21	1.00	1.00	0.10
alpha-BD	0.95	1.00	0.97	1.00	1.00	1.00
alpha-hmodal	1.00	1.00	0.71	1.00	1.00	0.55
alpha-FM	1.00	0.93	0.50	1.00	1.00	0.14
alpha-RP	1.00	0.93	0.50	1.00	1.00	0.14

Table 2: True positive average levels for 20% and 5% polluted data under all normally-distributed models. Citation of packages used can be found in Febrero-Bande and Oviedo de la Fuente (2012),Tarabelloni et al. (2018), Segaert et al. (2020), Ojo et al. (2020), Kosiorowski and Zawadzki (2020)

	Magnitude	Shape	Covariance	Mixed	Trend	Isolated
20% Outlyingness						
boxplot-MBD	1.16	1.02	1.02	2.00	1.96	1.00
boxplot-hmodal	1.16	1.02	1.02	2.00	1.99	1.01
boxplot-FM	1.16	1.02	1.02	2.00	1.98	1.00
boxplot-RP	1.10	1.00	1.01	1.98	1.90	1.00
centrality.stability	1.73	0.99	1.22	1.00	1.20	1.00
outliergram	1.55	1.00	1.04	1.00	1.00	1.00
outliergram_m	1.58	1.40	1.35	1.97	1.95	1.02
bagplot	0.66	1.00	0.99	1.00	1.00	1.00
hdr	0.58	0.62	0.88	0.66	0.64	1.00
directional	1.57	1.96	1.89	1.97	1.98	1.15
seq-O	1.87	1.94	1.74	1.96	1.97	1.00
seq-T-MBD	1.16	1.25	1.21	2.00	2.00	1.40
seq-T-hmodal	1.16	1.36	1.24	2.00	2.00	2.00
seq-T-FM	1.16	1.22	1.15	2.00	2.00	2.00
seq-T-RP	1.11	1.18	1.19	2.00	2.00	2.00
alpha-MBD	1.75	1.00	1.09	1.33	1.34	1.01
alpha-BD	1.57	1.80	1.75	1.78	1.79	1.43
alpha-hmodal	1.88	1.57	1.47	1.61	1.64	1.14
alpha-FM	1.91	1.16	1.29	1.62	1.47	1.01
alpha-RP	1.91	1.17	1.29	1.61	1.49	1.01
5% Outlyingness						
boxplot-MBD	1.34	1.07	1.05	2.00	2.00	1.00
boxplot-hmodal	1.33	1.07	1.05	2.00	2.00	1.23
boxplot-FM	1.30	1.04	1.03	2.00	2.00	1.00
boxplot-RP	1.27	1.03	1.03	2.00	1.99	1.00
centrality.stability	1.96	1.71	1.37	1.98	1.98	1.00
outliergram	1.59	1.66	1.21	1.97	1.98	1.00
outliergram_m	1.54	1.93	1.52	1.94	1.94	1.06
bagplot	0.61	1.00	0.99	1.00	1.00	1.00
hdr	0.53	0.50	0.86	0.50	0.50	1.00
directional	1.58	1.93	1.88	1.93	1.93	1.66
seq-O	1.90	1.90	1.79	1.91	1.90	1.13
seq-T-MBD	1.34	1.39	1.27	2.00	2.00	2.00
seq-T-hmodal	1.33	1.46	1.32	2.00	2.00	2.00
seq-T-FM	1.30	1.33	1.18	2.00	2.00	2.00
seq-T-RP	1.25	1.26	1.26	2.00	2.00	2.00
alpha-MBD	1.83	1.22	1.10	1.85	1.85	0.99
alpha-BD	1.85	1.73	1.73	1.73	1.73	1.74
alpha-hmodal	1.82	1.82	1.57	1.82	1.82	1.41
alpha-FM	1.85	1.75	1.36	1.82	1.82	1.00
alpha-RP	1.85	1.75	1.36	1.83	1.82	1.00

Table 3: $TP+1/FP+1$ ratio for 20% and 5% polluted data under all models

	Magnitude	Shape	Covariance	Mixed	Trend	Isolated
20% Outlyingness TP ratio						
boxplot-MBD	0.04	0.01	0.08	0.42	0.11	0.01
boxplot-hmodal	0.05	0.02	0.09	0.48	0.14	0.01
boxplot-FM	0.04	0.01	0.08	0.39	0.11	0.01
boxplot-RP	0.03	0.01	0.06	0.25	0.07	0.01
centrality.stability	0.52	0.05	0.22	0.03	0.00	0.05
outliergram	0.35	0.00	0.02	0.00	0.00	0.01
outliergram_m	0.51	0.11	0.35	0.53	0.51	0.05
bagplot	0.53	1.00	0.97	1.00	1.00	0.99
hdr	0.34	0.39	0.69	0.17	0.25	0.94
directional	0.53	0.99	0.90	1.00	1.00	0.08
seq-O	0.67	0.38	0.78	1.00	1.00	0.12
seq-T-MBD	0.06	0.03	0.48	0.94	0.12	0.07
seq-T-hmodal	0.06	0.05	0.53	0.99	0.17	1.00
seq-T-FM	0.05	0.03	0.42	0.93	0.12	0.29
seq-T-RP	0.05	0.03	0.46	0.90	0.08	0.59
alpha-MBD	0.47	0.08	0.16	0.13	0.08	0.10
alpha-BD	0.40	0.78	0.79	1.00	0.99	0.44
alpha-hmodal	0.81	0.34	0.59	0.58	0.25	0.16
alpha-FM	0.74	0.12	0.33	0.25	0.20	0.12
alpha-RP	0.73	0.13	0.34	0.25	0.20	0.12
$TP + 1/FP + 1$ ratio						
boxplot-MBD	1.03	1.01	1.07	1.40	1.10	1.00
boxplot-hmodal	1.04	1.01	1.08	1.47	1.13	1.00
boxplot-FM	1.03	1.01	1.07	1.38	1.10	1.00
boxplot-RP	1.02	1.00	1.06	1.24	1.07	1.00
centrality.stability	1.50	0.99	1.20	0.97	0.94	1.00
outliergram	1.35	1.00	1.02	1.00	1.00	1.00
outliergram_m	1.48	1.10	1.33	1.52	1.50	1.01
bagplot	0.77	1.00	0.99	1.00	1.00	1.00
hdr	0.69	0.71	0.86	0.60	0.63	1.00
directional	1.46	1.92	1.83	1.92	1.93	1.03
seq-O	1.61	1.27	1.68	1.89	1.91	1.00
seq-T-MBD	1.03	1.01	1.44	1.90	1.10	1.04
seq-T-hmodal	1.03	1.02	1.50	1.94	1.13	1.94
seq-T-FM	1.03	1.01	1.40	1.89	1.10	1.26
seq-T-RP	1.02	1.01	1.43	1.86	1.06	1.55
alpha-MBD	1.42	0.98	1.08	1.06	1.00	1.00
alpha-BD	1.34	1.69	1.70	1.85	1.86	1.34
alpha-hmodal	1.70	1.23	1.46	1.47	1.13	1.03
alpha-FM	1.68	1.02	1.24	1.18	1.11	1.00
alpha-RP	1.66	1.02	1.25	1.18	1.11	1.00

Table 4: TP and $TP + 1/FP + 1$ ratio for 20% polluted data under all t-distributed models

3 Application to PM2.5 pollution data

One of the common usages of multivariate functional outlier detection is sample robustification on supervised classification procedures Ieva and Paganoni (2017). In this section, the findings of section 2 are applied to fit that purpose on a real dataset of air pollution in Medellín, Colombia.

We consider hourly records of air pollution from four measuring stations over the Metropolitan Area of Aburrá Valley in Medellín, Colombia. Two of them are traffic-level stations and two of them trend stations. Data are taken from a publicly available source (SIATA, 2021) and their characteristics are summarised in 5.

Station code	Station name	Location	Type of station
CEN-TRAF	Tráfico Centro	Center	Traffic
ITA-CJUS	Casa de Justicia Itagüí	Southwest	Trend
SUR-TRAF	Tráfico Sur	South	Traffic
MED-VILL	Villahermosa	East	Trend

Table 5: Attributes of the stations analyzed

The data time span comprehends 670 multivariate functional observations, starting on 2019/01/01 and ending on 2020/10/31. Since each observation takes into account hourly data, those 670 daily observations are in fact 16080 discrete hourly points.

3.1 Missing value imputation

During data preprocessing, some missing data were found, mainly associated to failures in the measuring devices. We used the same validation criteria that was provided by SIATA (2019) to discard pre-defined atypical observations. The application of those criteria lead to an amount of missing data that ranges from 82 days with missing hourly points in ITA-CJUS to 209 in MED-VILL. As shown in table 6, most of the days had one missing hourly point and very few of them had missing a whole day. Summarizing, 1.487 points (or 11.46% out of 16.080) were missing.

Table 6 shows the quantity of days according to the amount of missing hours. As depicted, most of the missing values where found isolatedly, from 1 to 4 hours per day at all monitoring stations.

The presence of moderate Pearson's cross-sectional correlation among the four variables (see Table 7) suggests probable gains from multivariate imputation. To take advantage of the complete dataset, we implemented the imputation procedure developed by Junger and Ponce de Leon (2015), specifically tested on air pollution multivariate time series data [7].

3.2 Classification procedure

To test the effects of data robustification on the accuracy of a classification procedure, we implemented the DD-G Classifier, developed by Cuesta-Albertos, Febrero-Bande, and Oviedo de la Fuente (2017), based on Li, Cuesta-Albertos, and Liu (2012). Whilst in the traditional DD- classifier the method consists in applying classification methods to DD-Plots Liu et al. (1999), which are scatter plots where the dimensions are, respectively, depths of a point with respect to each variable of a bivariate data set, in the DD-G Classifier, depths are not calculated with respect to the distribution

[7] the imputation algorithm consists on the estimation of the smoothed mean value for each t using a cubic spline, followed by a modification of the EM algorithm made of three steps: 1. Replacing missing values by estimates, 2. estimate the parameters μ and Σ, 3. Estimate the level for each multivariate time series, 4. Re-estimate the missing values with new parameters (Junger & Ponce de Leon, 2015). The procedure was made using the R package mtsdi

	Number of days			
Missing hours	CEN-TRAF	ITA-CJUS	SUR-TRAF	MED-VILL
1	122	42	51	118
2	43	19	18	46
3	11	3	4	23
4	9	0	3	5
5	4	1	0	2
6	1	3	1	2
7	1	0	0	0
8	3	0	1	2
9	0	2	0	2
10	3	1	1	0
11	1	2	0	2
12	0	1	1	2
13	1	0	1	0
14	2	1	2	0
15	2	1	0	1
16	0	1	1	0
17	1	0	1	0
18	1	0	1	0
19	0	0	0	1
20	0	0	1	0
21	1	0	0	0
24	2	5	8	3
Total	208	82	95	209

Table 6: Number of days according to the number of missing values on each day, for days with missing values, per station.

	CEN-TRAF	ITA-CJUS	SUR-TRAF	MED-VILL
CEN-TRAF	1.00	0.75	0.69	0.77
ITA-CJUS	0.75	1.00	0.72	0.70
SUR-TRAF	0.69	0.72	1.00	0.62
MED-VILL	0.77	0.70	0.62	1.00

Table 7: Pearson's Correlation of four stations, omitting missing values.

of a specific variable but to the distribution of the multivariate (functional) subset of the sample grouped by a specific value of the dependent variable. From that perspective, we go from a multivariate functional dataset with a discrete response variable, to a g-variate dataset, each variable being the depth of a (multivariate functional) observation w.r.t group g_i, and a response variable with g levels.

In order to accurately predict the occurrence of non-critical episodes of possible health-injuring levels of PM2.5 contamination, we built as an identifier the 7-days forward-moving average of PM2.5 concentration. The indicator used for the classification task took the value of 1 if any of the four stations had 7-days-forward averages of the daily average concentration of PM2.5 pollution over 37 $\mu g/m^3$ (SIATA, 2019), which led us to dropping 7 of the 670 observations and keeping 663, out of which 542 are not critical and 121 (18.25%) are critical. Figure 3 shows the data with the classification criterium. Given that we are trying to classify 7-days-forward averages, the predictions in which there are outlying observations are prone to be moderated by those non-outlying and, hence, an improvement from robustification is hypotetically likely to happen.

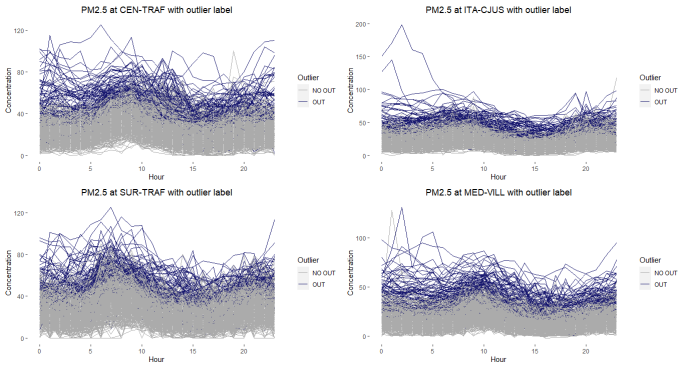


Fig. 3: Complete datasets with Identified outliers (blue).

	Models	Train.complete	Train.robust	Test.complete	Test.robust
1	knn	0.89	0.86	0.91	0.92
2	glm	0.90	0.87	0.92	0.92
3	np	0.90	0.87	0.92	0.91

Table 8: Accuracy rates of DD-G classifying with different classification procedures.

3.3 Results of outlier identification

Since outliers in this case are more likely magnitude outliers sith a high degree of dispersion, Alpha-hmodal identification criteria was used to robustify the sample, leading to the identification of 77 Multivariate Functional outlying observations, as can be seen in figure 3.

We used three classification methods inside the DD-G classifier methodology: K-nearest neighbours, logistic regression (glm) and non-parametric kernel regression. The methods were applied using the R package fda.usc(Febrero-Bande & Oviedo de la Fuente, 2012). To assess the accuracy of the methods, we split the sample into train (70%), robust train (same as train but without outliers), and test (not robustified) and calculated the usual accuracy index as the trace of the confusion matrix divided by the number of cases. The results of the classification procedure can be seen in table 8.

As seen, robustifying the samples had little or no effect on the classification algorithms. Both robustified and non-robustified methods had a classifying accuracy of 0.91 on the testing sample.

4 Conclusion

This article summarized methods for outlier identification on Multivariate Functional Data and showed the effect of robustification on the prediction of 7-day averages as critical periods of PM2.5 contamination on the Metropolitan Area of Medellín.

We were able to classify the best performing outlier identification mechanisms for low and high levels of pollution and for heavy-tailed distributions, finding that mixed, trend and isolated outliers under our simulation can be better identified

by sequential transformations, boxplot and directional outlyingness methods, and that magnitude, shape and covariance outliers are can be better identified by the directional outlyingness method, sequential transformations with the outlyingness method, alpha-trimming with hmodal depth and centrality-stability plots.

We showed that an undocumented procedure with low computational costs as the alpha-trimming with h-modal depth has a good performance on identifying magnitude outliers.

When verifying the effect of robustification on the classification accuracy, we showed that there is no clear effect of robustification. This can be due to two main reasons: the atypical nature of the indicator variables for environmental phenomena, that could be endogenous to the outlyingness of observations, and the fact that we are correlating many measuring stations on a single pollutant instead of many pollutants on a single station, which can lead to more correlated variables and, possibly, more diverse results.

The topic of Multivariate Functional Data Analysis of environmental data is a fertile field. Some future research starting from this methodology could include the evaluation of many more functions on the sequential transformations procedure, and the modification of data structure with other pollutants and stations. Also, further consideration of spatial dependence can also be of importance for future work.

References

Berrendero, J.R., Justel, A., Svarc, M. (2011). Principal components for multivariate functional data. *Computational Statistics and Data Analysis*, *55*(9), 2619–2634.

10.1016/j.csda.2011.03.011

Claeskens, G., Hubert, M., Slaets, L., Vakili, K. (2014). Multivariate functional halfspace depth. *Journal of the American Statistical Association*, *109*(505), 411–423.

10.1080/01621459.2013.856795

Cuesta-Albertos, J.A., Febrero-Bande, M., Oviedo de la Fuente, M. (2017). The DD G -classifier in the functional setting. *Test*, *26*(1), 119–142.

10.1007/s11749-016-0502-6

Cuesta-Albertos, J.A., & Nieto-Reyes, A. (2008). The random Tukey depth. *Computational Statistics and Data Analysis*, *52*(11), 4979–4988. https://arxiv.org/abs/0707.0167
10.1016/j.csda.2008.04.021

Cuevas, A., Febrero, M., Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics and Data Analysis*, *51*(2), 1063–1074.

10.1016/j.csda.2005.10.012

Cuevas, A., Febrero, M., Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, *22*(3), 481–496.

10.1007/s00180-007-0053-0

Dai, W., & Genton, M.G. (2019). Directional outlyingness for multivariate functional data. *Computational Statistics and Data Analysis*, *131*, 50–65. Retrieved from https://doi.org/10.1016/j.csda.2018.03.017 https://arxiv.org/abs/1612.04615
10.1016/j.csda.2018.03.017

Dai, W., Mrkvička, T., Sun, Y., Genton, M.G. (2020). Functional outlier detection and taxonomy by sequential transformations. *Computational Statistics and Data Analysis*, *149*(11901573). https://arxiv.org/abs/1808.05414

Febrero, M., Galeano, P., González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics*, *19*(4), 331–345.

10.1002/env.878

Febrero-Bande, M., Galeano, P., González-Manteiga, W. (2007). A functional analysis of NOx levels: location and scale estimation and outlier detection. *Reports in statistics and operations research*, *22*(3), 481–496.

10.1007/s00180-007-0053-0

Febrero-Bande, M., & Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, *51*(4), 1–28. Retrieved from http://www.jstatsoft.org/v51/i04/

Fraiman, R., & Muniz, G. (2001). Trimmed means for functional data. *Test*, *10*(2), 419–440.

10.1007/BF02595706

Hubert, M., Rousseeuw, P.J., Segaert, P. (2015). Multivariate functional outlier detection. *Statistical Methods and Applications*, *24*(2), 177–202.

10.1007/s10260-015-0297-8

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, *52*(12), 5186–5201.

10.1016/j.csda.2007.11.008

Hyndman, R.J., & Shang, H.L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, *19*(1), 29–45.

10.1198/jcgs.2009.08158

Ieva, F., & Paganoni, A.M. (2013). Depth measures for multivariate functional data. *Communications in Statistics - Theory and Methods*, *42*(7), 1265–1276.

10.1080/03610926.2012.746368

Ieva, F., & Paganoni, A.M. (2017). Component-wise outlier detection methods for robustifying multivariate functional samples. *Statistical Papers*, 1–20.

10.1007/s00362-017-0953-1

Ieva, F., Paganoni, A.M., Romo, J., Tarabelloni, N. (2019). Roahd package: Robust analysis of high dimensional data. *R Journal*, *11*(2), 291–307.

10.32614/RJ-2019-032

Junger, W.L., & Ponce de Leon, A. (2015). Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, *102*, 96–104. Retrieved from http://dx.doi.org/10.1016/j.atmosenv.2014.11.049

10.1016/j.atmosenv.2014.11.049

Kosiorowski, D., & Zawadzki, Z. (2020). Depthproc an r package for robust exploration of multidimensional economic phenomena [Computer software manual].

Li, J., Cuesta-Albertos, J.A., Liu, R.Y. (2012). DD-classifier: Nonparametric classification procedure based on DD-plot. *Journal of the American Statistical Association*, *107*(498), 737–753.

10.1080/01621459.2012.688462

Liang, D., Zhang, H., Chang, X., Huang, H. (2020). Modeling and Regionalization of China's PM2.5 Using Spatial-Functional Mixture Models.

Journal of the American Statistical Association, *0*(0), 1–70. Retrieved from https://doi.org/10.1080/01621459.2020.1764363

10.1080/01621459.2020.1764363

Liu, R.Y. (1990). On a notion of data depth based on random simplices. *Annals of Statistics*, *18*(1), 1403–405–414.

Liu, R.Y., Parelius, J.M., Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Annals of Statistics*, *27*(3), 783–858.

10.2307/120138

López-Pintado, S., & Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, *104*(486), 718–734.

10.1198/jasa.2009.0108

López-Pintado, S., Sun, Y., Lin, J.K., Genton, M.G. (2014). Simplicial band depth for multivariate functional data. *Advances in Data Analysis and Classification*, *8*(3), 321–338.

10.1007/s11634-014-0166-6

Martínez, J., Saavedra, Á., García-Nieto, P.J., Piñeiro, J.I., Iglesias, C., Taboada, J., ... Pastor, J. (2014). Air quality parameters outliers detection using functional data analysis in the Langreo urban area (Northern Spain). *Applied Mathematics and Computation*, *241*(2), 1–10.

10.1016/j.amc.2014.05.004

Nagy, S., Gijbels, I., Hlubinka, D. (2017). Depth-based recognition of shape outlying functions. *Journal of Computational and Graphical Statistics*, *26*(4), 883-893. Retrieved from https://doi.org/10.1080/10618600.2017.1336445 https://arxiv.org/abs/https://doi.org/10.1080/10618600.2017.1336445

10.1080/10618600.2017.1336445

Ojo, O.T., Lillo, R.E., Fernandez Anta, A. (2020). fdaoutlier: Outlier detection tools for functional data analysis [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=fdaoutlier (R package version 0.1.1)

Ramsay, J., & Silverman, B. (2005). *Functional Data Analysis*. Springer.

Rousseeuw, P.J., & Ruts, I. (1999). The Bagplot: A Bivariate Boxplot. *Statistical Computing and Graphics*, *53*(4), 382–387.

Sánchez-Lasheras, F., Ordóñez-Galán, C., García-Nieto, P.J., García-Gonzalo, E. (2020). Detection of outliers in pollutant emissions from the Soto de Ribera coal-fired power plant using functional data analysis: a case study in northern Spain. *Environmental Science and Pollution Research*, *27*(1), 8–20.

10.1007/s11356-019-04435-4

Segaert, P., Hubert, M., Rousseeuw, P., Raymaekers, J. (2020). mrfdepth: Depth measures in multivariate, regression and functional settings [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=mrfDepth (R package version 1.0.12)

Shaadan, N., Deni, S.M., Jemain, A.A. (2012). Assessing and comparing PM10 pollutant behaviour using functional data approach. *Sains Malaysiana*, *41*(11), 1335–1344.

Shaadan, N., Jemain, A.A., Latif, M.T., Deni, S.M. (2015). Anomaly detection and assessment of PM10 functional data at several locations in the Klang Valley, Malaysia. *Atmospheric Pollution Research*, *6*(2), 365–375. Retrieved from http://dx.doi.org/10.5094/APR.2015.040

10.5094/APR.2015.040

SIATA (2019). *Generalidades de la información Red de Calidad del Aire del Valle de Aburrá* (Tech. Rep.). Medellín: Área Metropolitana del Valle de Aburrá. Retrieved from https://siata.gov.co/descarga_siata/index.php/info/aire/

SIATA (2021). *Información de calidad del aire.* Retrieved from https://siata.gov.co/descarga_siata/index.php/index2/calidad_aire/

Sun, Y., & Genton, M.G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, *20*(2), 316–334.

10.1198/jcgs.2011.09224

Tarabelloni, N., Arribas-Gil, A., Ieva, F., Paganoni, A.M., Romo, J. (2018). roahd: Robust analysis of high dimensional data [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=roahd (R package version 1.4)

Torres, J.M., Nieto, P.J., Alejano, L., Reyes, A.N. (2011). Detection of outliers in gas emissions from urban areas using functional data analysis. *Journal of Hazardous Materials*, *186*(1), 144–149.

10.1016/j.jhazmat.2010.10.091

Torres, J.M., Pérez, J.P., Val, J.S., McNabola, A., Comesaña, M.M., Gallagher, J. (2020). A functional data analysis approach for the detection of air pollution episodes and outliers: A case study in Dublin, Ireland. *Mathematics*, *8*(2).

10.3390/math8020225

Wang, Y., Xu, K., Li, S. (2020). The functional spatio-temporal statistical model with application to O3 pollution in Beijing, China. *International Journal of Environmental Research and Public Health*, *17*(9).

10.3390/ijerph17093172

World Health Organization (2006). *WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide - Global Update 2005* (Tech. Rep.). Retrieved from https://apps.who.int/iris/bitstream/handle/10665/69477/WHO˙SDE˙PHE˙OEH˙06
10.1007/s12011-019-01864-7

Zuo, Y., & Serfling, R. (2000). General Notions of Statistical Depth Function. *Statistics*, *28*(2), 461–482.