Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels

Article *in* Environmetrics · June 2008

DOI: 10.1002/env.878

CITATIONS
140

READS
1,131

3 authors:

Manuel Febrero-Bande
University of Santiago de Compostela
78 PUBLICATIONS **1,963** CITATIONS

SEE PROFILE

Pedro Galeano
University Carlos III de Madrid
49 PUBLICATIONS **718** CITATIONS

SEE PROFILE

Wenceslao González-Manteiga
University of Santiago de Compostela
235 PUBLICATIONS **3,514** CITATIONS

SEE PROFILE

ENVIRONMETRICS
Environmetrics (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/env.878

Outlier detection in functional data by depth measures, with application to identify abnormal NO_x levels

Manuel Febrero, Pedro Galeano*,† and Wenceslao González-Manteiga

Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, Spain

SUMMARY

This paper analyzes outlier detection for functional data by means of functional depths, which measures the centrality of a given curve within a group of trajectories providing center-outward orderings of the set of curves. We give some insights of the usefulness of looking for outliers in functional datasets and propose a method based in depths for the functional outlier detection. The performance of the proposed procedure is analyzed by several Monte Carlo experiments. Finally, we illustrate the procedure by finding outliers in a dataset of NO_x (nitrogen oxides) emissions taken from a control station near an industrial area. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: depths; functional median; functional trimmed mean; nitrogen oxides; outliers; smoothed bootstrap

1. INTRODUCTION

In recent years, there exists an increasing interest on functional data analysis (FDA), specially in environmental, meteorological, medical, and economic contexts. FDA deals with the case in which the data are repeated measurements of the same subject densely taken over an ordered grid of points belonging to an interval of finite length. Thus, for each subject, we observe a function and, albeit the recording points are really discrete, we may regard the entire function as being continuously observed. For instance, Figure 1 shows the dataset of nitrogen oxides (NO_x) emission levels measured by an environmental control station close to an industrial area in Poblenou, a neighborhood in Barcelona. NO_x are one of the most important pollutants, precursors of ozone formation and contributors to global warming. NO_x is primarily caused by combustion processes in sources that burn fuels such as motor vehicles, electric utilities, and industries. The control station measures NO_x levels in $\mu g/m^3$ every hour of every day. The dataset starts on 23 February and ends on 26 June, in 2005. We split the whole sample of hourly measures in a dataset of functional trajectories of 24 h observations. Thus, each curve represents the evolution of the levels in 1 day. The figure shows the behavior of the NO_x levels, which increase in the morning, attain their largest values around 8:00am, then decrease until 14:00pm, and increase again at the evening. As the control station is located at the city center, there is an apparent

*Correspondence to: P. Galeano, Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, 15782 Santiago de Compostela, Spain.
†E-mail: pgaleano@usc.es

Received 24 October 2006
Accepted 22 June 2007

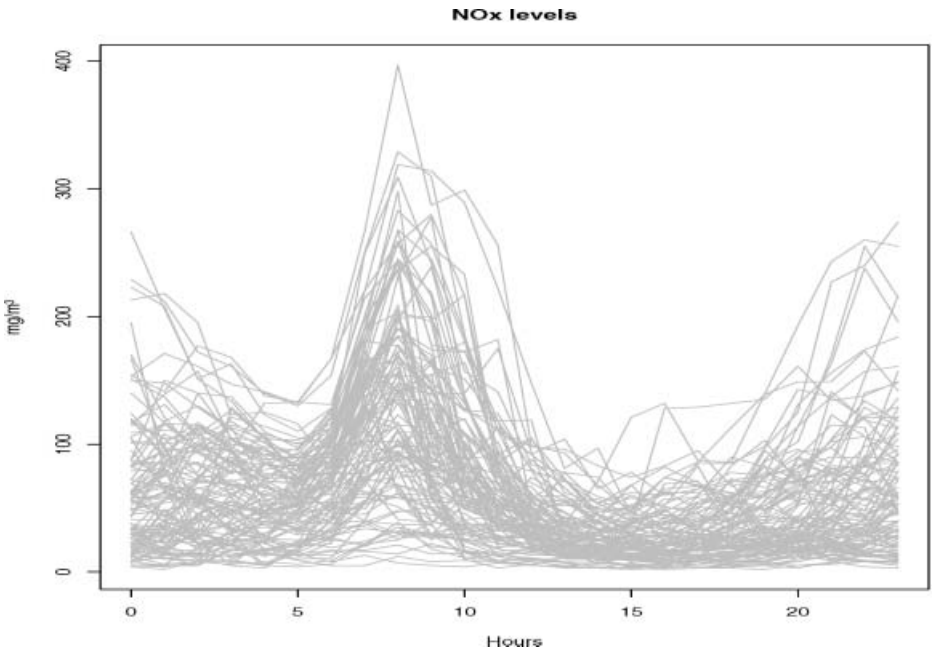

Figure 1. NO_x levels measured by a control station in Poblenou. Each curve represents a day

large influence of traffic on the measured NO_x levels. It is specially meaningful that several curves attain extreme values at some of the observed time points. For instance, the curve that attains around $400\,\mu g/m^3$ at 8:00am is abnormally large compared with the rest of values. It is important to identify days or periods in which the NO_x levels are significatively large, because these outlying observations may allow us to find out sources which produce abnormally large NO_x emissions.

The analysis of outliers is an important aspect of any statistical analysis of data, see, for instance, Barnett and Lewis (1994) for a general review on the topic. Although the presence of outliers may have significative impact on statistical methodology in many different ways, their analysis in functional data has been seldom addressed. For instance, robust estimates of the center of the functional distribution such as the functional trimmed means were introduced by Fraiman and Muniz (2001). The main aim of this paper is to analyze outliers in functional data by providing a definition of what an outlier is in these settings and by proposing a nonparametric outlier detection procedure for functional outlier detection based on the concept of functional depth.

The first possibility to look for outliers in functional samples can be the application of a multivariate outlier detection procedure to functional samples. Unfortunately, there are several reasons why multivariate statistical methods are not well suited for functional data. First, functional data are observations of a smooth random process observed at discrete time points. Thus, the time correlation structure is ignored when using multivariate statistical methods. Second, the infinite-dimensional nature of functional variation makes that in many situations, the number of grid points is larger than the number of subjects. It is well known that the most usual multivariate statistical methods are affected by the curse of dimensionality, i.e., these methods are not able to deal with situations in which the number of variables is larger than the number of individuals in the sample. Third, few distributional

DOI: 10.1002/env

assumptions are usually imposed in analyzing functional datasets. On the contrary, multivariate statistical methods are implicitly restricted to Gaussian or elliptical populations. In functional data, the covariance and correlation structures are difficult to interpret, as they do not usually give a fully comprehensible presentation of the structure of the variability in the observed data. For all of these reasons, a lot of statistical methods have been developed for FDA extending those of multivariate data analysis to the functional framework. For instance, and among others, Silverman (1996), Boente and Fraiman (2000), and Hall and Hosseini-Nasab (2006) studied functional principal component analysis (FPCA). Cardot *et al*. (1999), Cuevas *et al*. (2002), and Cai and Hall (2006) analyzed regression with functional regressors and real and/or functional responses. Ferraty and Vieu (2003) and Cuesta-Albertos and Fraiman (2007) considered the classification and discrimination of random curves, and Cuevas *et al*. (2004) developed the functional analysis of variance. Monographs on FDA are Ramsay and Silverman (2004, 2005), which provide a large catalog of methods and case studies for handling functional data, and Ferraty and Vieu (2006), which present a nonparametric approach for analyzing functional samples.

In order to look for outliers in functional data, the use of univariate and multivariate outlier detection methods to functional samples encounter several additional drawbacks to those commented previously. For instance, a procedure which detects univariate outliers in the sets of curve values at each observed time point may be inadequate because the curve values can be nonoutlying at each time point but the curve be a functional outlier. Also, multivariate outlier detection methods are not well suited in functional settings for several reasons. First, plotting methods are not appropriate for dimensions higher than 3, which is never the case in functional data. Second, outlier detection methods for multivariate samples are highly affected by the curse of dimensionality. This is also a drawback for robust methods. Third, outlier detection and robust methods for multivariate samples are an example of multivariate procedures implicitly restricted to multivariate normal or elliptical samples, as they are usually based on robust estimates of a location vector and a scatter matrix of the data. These are distributional assumptions very restrictive for functional data. Moreover, these methods are, more or less, based on the use of robustified versions of the Mahalanobis distance. Thus, the regions with outliers have elliptical contours which are inadequate from a functional point of view. In summary, approaches of a functional nature are needed. The procedure proposed here, based on functional depths, is of a functional nature and it is not affected by the previous drawbacks. The aim of functional depths is to measure the centrality of a given curve within a group of trajectories. Thus, depths provide a center-outward ordering of the set of curves. In fact, the curve with maximum depth may be defined as an estimate of the center of the functional distribution. Therefore, depth and outlyingness are inverse notions, and functional outliers, which are curves that are expected to be far away from the center of the data, will correspond to curves of significatively low depth. Consequently, a simple way to look for functional outliers is to detect which curves have a significant low depth. Also, this is a nonparametric approach which fits well with the absence of model structural assumptions in the data.

The rest of this paper is organized as follows. Section 2 reviews depth measures for functional data. Section 3 studies outliers in functional data and proposes an outlier detection procedure based on functional depths. Section 4 analyzes the performance of the proposed procedure by means of several Monte Carlo experiments. Finally, Section 5 illustrates the proposed procedure with the NO_x data, in which several functional outliers are found, showing the strong incidence of traffic on the NO_x emission levels.

2. DEPTH MEASURES FOR FUNCTIONAL DATA

Depth measures were originally introduced in multivariate data analysis for measuring the centrality of a point $x \in R^d$, with respect to a data cloud generated from F, a probability distribution in R^d. Thus,

DOI: 10.1002/env

depths provide a way to order points in the Euclidean space from center to outward, such that points near the center should have higher depth and points far from the center should have lower depth. Several data depths have been proposed in the literature. Some well-known examples are the halfspace depth (Tukey, 1975), the Oja's depth (Oja, 1983), the simplicial depth (Liu, 1990), and the projection depth (Zuo, 2003). See Zuo and Serfling (2000) for an extensive analysis of the definitions, properties, and applications of multivariate depths. For instance, these authors found that the halfspace depth behaves very well overall in comparison with various competitors based in properties such as affine invariance, maximality at center, monotonicity relative to the deepest point, and vanishing at infinity.

Although multivariate depths have been derived from several statistical notions and have quite different definitions, depths are very similar in the case of continuous one-dimensional random variables. For instance, if $x_1, \ldots, x_n$ is a sample drawn from a random variable with cumulative distribution function F, the halfspace and the simplicial depths of x_i with respect to the sample $x_1, \ldots, x_n$, are given by

$$\mathrm{HD}_n(x_i) = \min\{F_n(x_i), 1 - F_n(x_i)\},$$

and,

$$\mathrm{SD}_n(x_i) = 2F_n(x_i)(1 - F_n(x_i)),$$

respectively, where F_n is the empirical cumulative distribution function of the sample $x_1, \ldots, x_n$. Both depths provide very similar ordered samples $x^{(1)}, \ldots, x^{(n)}$, such that $x^{(1)}$ is the deepest point and $x^{(n)}$ is the less deepest one.

Recently, the notion of depth has been extended to functional data by Fraiman and Muniz (2001) and Cuevas *et al*. (2006, 2007) (see also, López-Pintado and Romo (2006), for an alternative point of view based on the graphic representation of curves). The aim of data depth for functional data is to measure the centrality of a given curve, x_i, within a set of curves, $x_1, \ldots, x_n$, generated from a stochastic process $X(\cdot)$ with sample paths in $C([a, b])$, the space of continuous functions defined on the interval $[a, b] \subset R$. In other words, the idea under a functional depth is to measure how long a curve remains in the middle of the whole group of trajectories. We briefly review three functional depths that will be posteriorly used for outlier detection.

1. *The Fraiman and Muniz depth*: Fraiman and Muniz (2001) were the first to introduce a functional data depth. Let $F_{n,t}(x_i(t))$ be the empirical cumulative distribution function of the values of the curves $x_1(t), \ldots, x_n(t)$ at a given time point $t \in [a, b]$, given by

$$F_{n,t}(x_i(t)) = \frac{1}{n}\sum_{k=1}^{n} I(x_k(t) \leq x_i(t)).$$

   where $I(\cdot)$ is an indicator function. Then, the Fraiman and Muniz functional depth, hereafter FMD, of a curve x_i with respect the set $x_1, \ldots, x_n$ is given by

$$\mathrm{FMD}_n(x_i) = \int_a^b D_n(x_i(t))\,\mathrm{d}t, \tag{1}$$

DOI: 10.1002/env

where $D_n(x_i(t))$ is the univariate depth of the point $x_i(t)$ given by

$$D_n(x_i(t)) = 1 - \left| \frac{1}{2} - F_{n,t}(x_i(t)) \right|.$$

Therefore, Equation (1) reduces to

$$\mathrm{FMD}_n(x_i) = \int_a^b \left[1 - \left| \frac{1}{2} - F_{n,t}(x_i(t)) \right| \right] \mathrm{d}t. \tag{2}$$

In many cases, the curves are observed at a discretized set of different time points $a \leq t_1 < \cdots < t_m \leq b$, and, consequently, a functional dataset of n identically distributed functional curves, $x_1, \ldots, x_n$, observed at a grid of points, $t_1, \ldots, t_m$, is given by

$$\{x_i(t_j);\ i = 1, \ldots, n;\ j = 1, \ldots, m\}. \tag{3}$$

If this is not the case, note that we can always use some interpolation method to obtain a transformed functional dataset observed at the same grid of points. See Ferraty and Vieu (2006) for more details on this approach. The sample FMD of the curves (Equation (3)) are obtained after approximating the integral (Equation (2)), for instance, by simply using Riemann sums, as follows:

$$\mathrm{SFMD}_n(x_i) = \sum_{j=2}^{m} \Delta_j \left[1 - \left| \frac{1}{2} - F_{n,t_j}(x_i(t_j)) \right| \right], \quad i = 1, \ldots, n \tag{4}$$

where $\Delta_j = (t_j - t_{j-1})$ and $F_{n,t_j}(x_i(t_j))$ is the empirical cumulative distribution function of the points $x_1(t_j), \ldots, x_n(t_j)$. Consequently, the sample FMD in Equation (4) can be written as

$$\mathrm{SFMD}_n(x_i) = \sum_{j=2}^{m} \Delta_j \left[1 - \left| \frac{1}{2} - \frac{1}{n} \sum_{k=1}^{n} I(x_k(t_j) \leq x_i(t_j)) \right| \right],$$

and, thus, the SFMD of the curve $x_i(t)$ reduces to a weighted sum of the univariate depths of the m sets $x_1(t_j), \ldots, x_n(t_j)$, for $j = 1, \ldots, m$.

2. *The h-modal depth*: Cuevas *et al*. (2006) introduced a functional depth based on the concept of mode. These authors defined a functional mode as the curve most densely surrounded by the rest of curves of the dataset. More precisely, the h-modal functional depth, hereafter MD, of a curve x_i, with respect the set of curves $x_1, \ldots, x_n$, is given by

$$\mathrm{MD}_n(x_i, h) = \sum_{k=1}^{n} K\left(\frac{\|x_i - x_k\|}{h} \right), \tag{5}$$

where $\| \ \|$ is a norm in the functional space, $K : R^+ \to R^+$ is a kernel function, and h is a bandwidth. Then, the functional mode is defined as the curve that attains the maximum value of Equation (5).

DOI: 10.1002/env

From the practical point of view, a functional norm, a kernel function, and a bandwidth h have to be chosen. Cuevas *et al*. (2006) recommended to use the L^2 and L^∞ norms, given by

$$\|x_i - x_k\|_2 = \left(\int_a^b (x_i(t) - x_k(t))^2 \, \mathrm{d}t \right)^{\frac{1}{2}} \qquad \|x_i - x_k\|_\infty = \sup_{t \in (a,b)} |x_i(t) - x_k(t)|,$$

and the truncated Gaussian kernel:

$$K(t) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right), \quad t > 0.$$

The bandwidth taken is the 15th percentile of the empirical distribution of $\{\|x_i - x_k\|,\ i, k = 1, \ldots, n\}$. Note that the bandwidth h does not need to provide a very good fit of the density because we are not directly concerned here with density estimation but with support estimation. Thus, the interest is in the values around the center of the distribution, which are not very sensitive to the choice of the bandwidth. In fact, a wide range of values of the bandwidth are appropriate with the only requirement that the bandwidth is not very small. See Cuevas and Fraiman (1997) and Cuevas *et al*. (2001) for more information on the choice of the bandwidth in these situations. If the curves are observed at discretized points, the functional norms are replaced by their empirical versions

$$\|x_i - x_k\|_2 = \left(\sum_{j=2}^{m} \Delta_j (x_i(t_j) - x_k(t_j))^2 \right)^{\frac{1}{2}} \qquad \|x_i - x_k\|_\infty = \sup_{j=1,\ldots,m} \left| x_i\left(t_j\right) - x_k(t_j) \right|.$$

3. *The Random projection depth*: Cuevas *et al*. (2007) considered the random projection depth based on measuring the depth of the functional data under projections and taking additional information of their derivatives. The basic idea is to project each functional curve and its first derivative along a random direction, defining a point in R^2. Now, a data depth in R^2 provides an order of the projected points. If we use a large number of random projections, the mean value of the depths of the projected points defines a depth for functional data. See Cuevas *et al*. (2007) for a detailed exposition of the random projection depth. Given the set of curves $x_1, \ldots, x_n$ and a direction v that belongs to an independent direction process $V(\cdot)$, $T_{i,v} = \langle v, x_i \rangle$ is the projection of x_i along the direction v, i.e.:

$$T_{i,v} = \langle v, x_i \rangle = \int_a^b v(t) x_i(t) \, \mathrm{d}t.$$

Similarly, $T'_{i,v} = \langle v, x'_i \rangle$ is the projection of the derivative of x_i, x'_i, along the direction v. Therefore, the pair $(T_{i,v}, T'_{i,v})$ is a point in R^2. Now, if $v_1, \ldots, v_P$ are P independent random directions, the random projection depth of a curve x_i is defined by

$$\mathrm{RPD}_n(x_i) = \frac{1}{P} \sum_{p=1}^{P} D_n\left(\left(\left\langle v_p, x_i \right\rangle, \left\langle v_p, x'_i \right\rangle \right), h \right), \tag{6}$$

DOI: 10.1002/env

where D_n is any depth defined in R^2 of the point $(\langle v_p, x_i\rangle, \langle v_p, x_i'\rangle) \in R^2$. In the following equation we take $D_n = \mathrm{MD}_n$, the h-modal depth. The sample version of the random projection depth has the same expression as that of Equation (6) but the inner product is given by

$$\langle v, x_i\rangle = \sum_{j=2}^{m} \Delta_j v(t_j) x_i(t_j).$$

The random directions can be generated by a Gaussian process defined in the interval $C([a, b])$.

Let $D_n(x_1), \ldots, D_n(x_n)$ be the functional depths of the set of trajectories $x_1, \ldots, x_n$ with any of the three functional depths introduced before. According to the multivariate case, the largest the value of $D_n(x_i)$ the deepest is the curve x_i, among the set of curves $x_1, \ldots, x_n$. Once the curves are ranked according to decreasing values of their depths, we get the ordered curves $x^{(1)}, \ldots, x^{(n)}$, such that $x^{(1)}$ is the deepest curve and $x^{(n)}$ is the less deepest one. This rank allows to define location estimates for functional data. For instance, as seen before, a functional mode is defined as the curve that attains the maximum value of the MD depth in Equation (5). Also, Fraiman and Muniz (2001) defined functional trimmed means as the average of the most deepest $n - [\alpha n]$ curves with the FMD depth in Equation (1), as follows:

$$\mathrm{FTM}_\alpha(x_1, \ldots, x_n) = \frac{1}{n - [\alpha n]} \sum_{i=1}^{n-[\alpha n]} x^{(i)}, \tag{7}$$

where α is such that $0 \leq \alpha \leq (n-1)/n$ and $[\cdot]$ denotes the integer part. If only one curve achieves the depth maximum value, a functional median is also defined as the functional trimmed mean (Equation (7)) with $\alpha = (n-1)/n$ and coincides with the deepest curve:

$$\mathrm{FMED}(x_1, \ldots, x_n) = x^{(1)}.$$

Elsewhere, the functional median is defined as the average of those curves which maximizes the depth. Note that the functional trimmed mean taking $\alpha = 0$ is the functional mean.

3. OUTLIERS IN FUNCTIONAL DATA

Outliers in a functional dataset can arise for, at least, two reasons. First, outliers may be curves with gross errors such as measurement, recording, and typing mistakes. These errors should be identified and corrected whenever possible. Second, outliers may be real data curves in the sense that they are not gross errors but are somehow suspicious or surprising as they do not follow the same pattern as that of the rest of curves. We are interested in detecting and examining precisely such surprising curves as, first, they may bias our functional estimates and we would like to prevent this, and, second, they may allow to discover which sources produce these outlying curves.

A rigorous definition of outlier in functional settings has not been given. In what follows, we consider that a curve is an outlier if it has been generated by a stochastic process with a different distribution than the rest of curves, which are assumed to be identically distributed. Therefore, we assume that the whole set of curves has been drawn from the same stochastic process, and curves that are not compatible with

DOI: 10.1002/env

this assumption are outliers. We point out that this definition covers most of the possible contaminants models. For instance, a curve could be an outlier if it is significatively far away from the expected function of the stochastic process or have a different shape than the rest of curves. Note also that this definition includes the case of curves which are different from the rest only during some subintervals of the whole observation period.

In order to identify outliers in functional datasets, we make use of functional depths. As mentioned in the introduction, depth and outlyingness are inverse notions, so that if an outlier is in the dataset, the corresponding curve will have a significant low depth. Therefore, a way to detect the presence of functional outliers is to look for curves with lower depths. Consequently, we propose the following functional outlier detection procedure for detecting outliers in a given dataset of functional curves $x_1, \ldots, x_n$:

(1) Obtain the functional depths $D_n(x_1), \ldots, D_n(x_n)$, for one of the functional depths, FMD, MD, or RPD.
(2) Let $x_{i_1}, \ldots, x_{i_k}$ be the k curves such that $D_n(x_{i_k}) \leq C$, for a given cutoff C. Then, assume that $x_{i_1}, \ldots, x_{i_k}$ are outliers and delete them from the sample.
(3) Then, come back to step 1 with the new dataset after deleting the outliers found in step 2. Repeat this until no more outliers are found.

Some remarks on the proposed procedure are in order. First, the idea behind the procedure is to identify functional outliers as the curves whose depths are lower outliers under the distribution of the functional depth considered (FMD, MD, and RPD). We will make use of this idea for obtaining the cutoff C. Second, the main contribution of the procedure is to point out curves that need further attention. Thus, once a set of outliers has been detected, it is compulsory to look for the reasons why these trajectories have a different behavior than the rest. Third, step 3 of the procedure is introduced to avoid masking effects. Masking appears when true outliers mask the presence of others. Therefore, if a set of outliers is masked in one iteration, they may be revealed in a later iteration after removing detected outliers.

Obviously, the key point of the previous algorithm is to determinate the cutoff C, which has to be chosen to ensure a reasonable level of type I errors. We select C such that, in the absence of outliers, the percentage of correct observations mislabeled as outliers is approximately equal to 1%. In other words, we select C such that, in the absence of outliers:

$$\Pr(D_n(x_i) \leq C) = 0.01, \quad i = 1, \ldots, n.$$

Thus, the C taken is the 1th percentile of the distribution of the functional depth under consideration. Unfortunately, the distribution of functional depths is unknown and appears to be hard to derive in a general setting. Thus, the cutoff C in our procedure is found by estimating this percentile, making use of the observed sample curves. As the sample may be contaminated by outliers, we have to be able to obtain a robust estimate of the percentile based on the sample. For that, we make use of two alternative bootstrap procedures. The first one is based on trimming the sample of suspicious curves, obtain smoothed bootstrap sets from the trimmed sample, and estimate the percentile based on the bootstrap sets. The smoothed bootstrap procedure based on trimming runs as follows:

(1) Obtain the functional depths $D_n(x_1), \ldots, D_n(x_n)$, for one of the functional depths, FMD, MD, or RPD.
(2) Obtain B standard bootstrap samples of size n from the dataset of curves obtained after deleting the α% less deepest curves. The bootstrap samples are denoted by x_i^b, for $i = 1, \ldots, n$ and $b = 1, \ldots, B$.

DOI: 10.1002/env

(3) Obtain smoothed bootstrap samples $y_i^b = x_i^b + z_i^b$, where z_i^b is such that $\left(z_i^b(t_1), \ldots, z_i^b(t_m)\right)$ is normally distributed with mean 0 and covariance matrix $\gamma\Sigma_x$, where Σ_x is the covariance matrix of $x(t_1), \ldots, x(t_m)$ and γ is a bootstrap smoothing parameter. Let y_i^b, $i = 1, \ldots, n$ and $b = 1, \ldots, B$ be these samples.
(4) For each bootstrap set $b = 1, \ldots, B$, obtain C^b as the empirical 1% percentile of the distribution of the depths, $D\left(y_i^b\right)$, $i = 1, \ldots, n$.
(5) Take C as the median of the values of C^b, $b = 1, \ldots, B$.

The second procedure for estimating the cutoff C is based on bootstrapping the curves of the original dataset with probability proportional to their depth. Thus, we sample more frequently the deepest points trying to avoid the presence of outliers in the bootstrap samples. The smoothed bootstrap procedure based on weighting runs as follows:

(1) Obtain the functional depths $D_n(x_1), \ldots, D_n(x_n)$, for one of the functional depths, FMD, MD, or RPD.
(2) Obtain B standard bootstrap samples from the curves in which each original curve is sampled with a probability proportional to its depth. The samples are denoted by x_i^b, for $i = 1, \ldots, n$ and $b = 1, \ldots, B$. Then, obtain smoothed bootstrap samples as in the step 3 of the first bootstrap procedure. Let y_i^b, $i = 1, \ldots, n$ and $b = 1, \ldots, B$, be these samples.
(3) For each bootstrap, set $b = 1, \ldots, B$, obtain C^b as the 1th empirical percentile of the distribution of the depths, $D\left(y_i^b(t)\right)$, $i = 1, \ldots, n$.
(4) Take C as the median of the values of C^b, $b = 1, \ldots, B$.

Some comments on these bootstrap procedures for determining the cutoff C are the following. First, if the number of outliers detected by the procedure exceeds substantially the significance level, it may be because the smoothed bootstrap procedures have given upward bias estimates of the 1th percentile of the distribution of the functional depths. Then, maybe a smallest significant level is needed. Second, the first bootstrap procedure depends on the trimmed sample used. The level α used can be chosen as the proportion of suspicious outliers in the sample. Third, both procedures depend on the smoothed parameter γ. The choice $\gamma = 0.05$ appears to work well in practice. Fourth, the cutoff C is the same for all the iterations of the procedure. One possibility may be to recalculate the cutoff C in each iteration using the bootstrap procedures described before, but this is discarded for two reasons. First, the computational cost increases with the obtention of a new cutoff. Second, the C given by the proposed procedures can be seen as robustified estimates of the 1th percentile of the distribution of the functional depths. If after one iteration no outliers remain in the dataset, there is a small probability of detecting a false outlier and, indeed this is the case as checked by simulations. Thus, in doing so, the procedure increases its type I error. Finally, the effectiveness of the proposed procedures is analyzed in the next section.

4. SIMULATION RESULTS

In this section, we explore several aspects of the outlier detection procedure for functional data introduced in Section 3 via a simulation study. The simulation results in this section and the analysis of the real data example in the next one have been carried out by means of various routines written by the authors in R (http://www.r-project.org/), which are available upon request. We consider functional data $x_1, \ldots, x_n$ obtained as realizations from a stochastic process $X(\cdot)$, which have continuous trajectories on the observation period $[a, b] = [0, 1]$. The mechanism model for generate curves without outliers is a

DOI: 10.1002/env

Gaussian process $X(t)$ of the form:

$$X(t) = E(t) + e(t),$$

where $E(t) = E(X(t)) = 30t(1-t)^{\frac{3}{2}}$ and $e(t)$ is a Gaussian process with mean 0 and covariance matrix:

$$E\left[e\left(t_i\right) e\left(t_j\right)\right] = 0.3 \exp\left(-\frac{\left|t_i - t_j\right|}{0.3}\right). \tag{8}$$

We also consider the alternative model for generating outlier curves, in which the mean function $E(t) = 30t(1-t)^{\frac{3}{2}}$ is replaced with $E(t) = 30t^{\frac{3}{2}}(1-t)$.

First, we study the type I error of the proposed procedure. For that, we generate 100 datasets with curves from the noncontaminated model and two sample sizes $n = 50$ and 100. The curves are observed at equidistant points $t_1 = 0, t_j = t_{j-1} + h_m$, $j = 2, \ldots, m$, where $h_m = (t_m - t_1)/(m-1)$. The number of grid points is $m = 30$. Then, we apply the proposed procedure to each one of the generated outliers-free datasets using the two cutoffs estimated by the smoothed bootstrap procedures based on trimming and weighting. For both procedures, we take $B = 200$ bootstrap samples and for the trimming procedure we use $\alpha = 0.1$. Thus, we delete the 10% less deepest curves. The results are shown in Table 1, where rows 3–5 report the mean percentage of false outliers detected by the procedure in each dataset. The type I error percentages are quite close to the nominal 1% in the two sample sizes considered. Note that when n increases the type I error percentages are closer to the nominal level.

Next, we consider the case of contaminated curves and make a simulation experiment in order to study the frequency detection of outliers of the proposed procedure. For that, we generate 100 datasets for sample sizes $n = 50$ and 100, where n_1 curves are generated from the noncontaminated model and n_0 curves are generated from the alternative model, such that for each sample size, $n_1 + n_0 = n$. More precisely, for sample size $n = 50$, we consider the pairs $n_1 = 49$ and $n_0 = 1$, $n_1 = 48$ and $n_0 = 2$, and $n_1 = 47$ and $n_0 = 3$, while for sample size $n = 100$, we consider the pairs $n_1 = 99$ and $n_0 = 1$, $n_1 = 98$ and $n_0 = 2$, and $n_1 = 97$ and $n_0 = 3$. The curves are observed at equidistant points with 30 grid points, as before. Then, we apply the proposed procedure to each one of the contaminated datasets generated with the two cutoffs estimated by using the smoothed bootstrap procedures based on trimming and weighting. As before, we take $B = 200$ and $\alpha = 0.1$.

The results are shown in Table 2, where rows 4–6 report the correct frequency detection of the procedure and rows 7–9 report the percentage of false outlier detection. The correct frequency detection is the number of times in which the procedure correctly detects the true outliers generated over the 100 generated datasets. The percentage of false outlier detection is the percentage of false outliers detected by the procedure in each dataset. Note that in order to evaluate the performance of the procedure, we have to take into account the trade off between frequency detection and false outlier detection rate. Taking

Table 1. Percentage of Type I errors with outliers free datasets for $n = 50$ and 100

	$n = 50$		$n = 100$	
	Trimming	Weighting	Trimming	Weighting
FMD	1.60	1.20	1.17	1.03
MD	1.70	1.44	0.99	1.01
RPD	1.42	1.10	1.36	1.22

DOI: 10.1002/env

Table 2. Correct frequency detection (up) and false outliers rate (down) of the proposed procedure for detecting functional outliers, where n is the sample size, n_0 denotes the number of outliers, Tm. and Wg. stand for trimming and weighting, respectively, and FMD, MD, and RPD denote the Fraiman and Muniz depth, the h-modal depth, and the random projection depth, respectively

	$n_0 = 1$				$n_0 = 2$				$n_0 = 3$			
	$n = 50$		$n = 100$		$n = 50$		$n = 100$		$n = 50$		$n = 100$	
	Tm.	Wg.	Tm.	Wg.	Tm.	Wg.	Tm.	Wg.	Tm.	Wg.	Tm.	Wg.
Frequency detection												
FMD	66	67	64	63	69	13	59	23	58	1	36	2
MD	98	100	100	100	100	79	100	83	99	63	97	57
RPD	100	89	100	100	100	46	97	66	97	25	100	37
False outliers												
FMD	1.06	0.54	0.96	0.60	1.16	0.40	0.97	0.39	1.10	0.42	1.01	0.39
MD	1.88	0.04	1.11	0.04	1.70	0.04	1.10	0.02	1.82	0.04	1.21	0.02
RPD	2.74	0.08	1.65	0.06	0.58	0.04	2.56	0.03	3.82	0.05	3.90	0.04

this into account, several conclusions arise from Table 2. First, the trimming version of the procedure has generally a larger frequency detection than the weighting version but at the cost of having a larger false outlier detection rate, usually slightly larger than the nominal 1%. Second, there are evident differences between the considered depths. In terms of correct frequency detection, the FMD have in general the worst performance, specially when the number of outliers increases. In this sense, the MD is slightly superior to the RPD. In terms of false outlier detection rate, there is not a general rule. Sometimes the RPD has the larger rate (see, for instance, the case of $n = 50$, $n_0 = 1$, and trimming cutoff) and sometimes the RPD has the smaller rate (see, for instance, the case of $n = 50$, $n_0 = 1$, and weighting cutoff). But when the number of outliers increases, the MD appears to have less false detection rate. Note also that the contaminant model here does not provide much information on the derivatives and, therefore, it is not very advantageous for the RPD, which makes use of this information. However, the RPD has a notable behavior under disadvantageous conditions. Third, the false outlier detection rates appear to depend only on the bootstrap procedure of selecting the cutoff, but do not depend on the sample size considered. Fourth, the computational cost of the procedure depends on the depth used. For instance, with the FMD, MD, and RPD depths, the procedure takes 13, 31, and 885 s respectively, in order to look for outliers in the same dataset. We note that the RPD has a much larger computational cost compared with the FMD and MD, because we are using $P = 50$ random projections. A smaller number of projections also provides close results and the computational cost is greatly reduced.

In summary, the proposed procedure has a good performance with several cases over the 90% of frequency detection. The MD and the RPD depths appear to have a better performance than the FMD depth. The trimming bootstrap estimate of the cutoff makes the procedure have a larger empirical power but at the cost of having a large false outlier detection rate than the weighting bootstrap estimate. Finally, the computational cost needed to detect outliers with the RPD allows to advice us on the use of the MD for outlier detection using the proposed procedure.

5. DETECTING OUTLIERS ON THE NO_x LEVELS

Finally, we illustrate the proposed procedure by looking for outliers in the dataset of NO_x levels described in Section 1. As described in Section 1, the data that we have at hand contain the NO_x levels

DOI: 10.1002/env

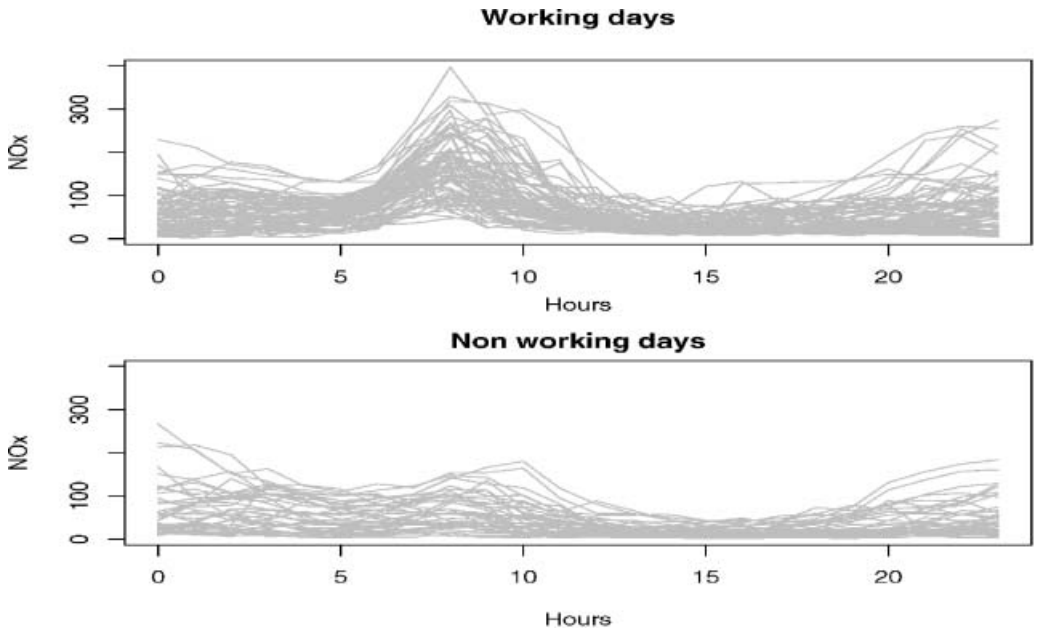

Figure 2. NO_x levels split into two groups: (up) working days; (down) nonworking days

measured in $\mu g/m^3$ taken every hour of every day starting on 23 February, and finishing on 26 June, in 2005. The whole hourly sample has been split to have a dataset of daily functional trajectories of 24 h observations. Unfortunately, complete data were available only for 115 days of the analysis because some of the measures are missing for several consecutive hours of some days. For simplicity, the days with incomplete periods have been discarded.

As it is expected, a great influence of the traffic on the measured NO_x levels was observed, the whole dataset was split into two groups formed by: (1) working days, which are the weekdays, and (2) nonworking days, which are the Saturdays, Sundays, and festive days. Figure 2 shows the 115 observed curves divided into two groups: working days (up) and nonworking days (down). As expected, there are apparent differences between both groups of curves, which include the presence of a few trajectories that are in some way different from the rest. In order to determinate if there are significant differences between the levels depending on the day of the week, we apply the anova test for functional data proposed by Cuevas *et al*. (2004), which is based on a statistic for which its asymptotic distribution can be approximated through a Monte Carlo procedure. In our case, we compute the anova test statistic to analyze differences between the two groups formed by 76 and 39 curves, respectively, assuming heteroscedasticity. The test strongly rejects the null hypothesis of common mean, so that we conclude that there are significant differences between the working and nonworking days. Consequently, we apply the procedure for the outlier detection of both groups separately.

We first apply the outlier detection procedure for the working days with the three different depths introduced in Section 3 and the two smoothed bootstrap cutoffs. The results are shown in Table 3, where rows 4–6 and 7 and 8 show the outliers detected by the procedure with the three depths for the working and nonworking days, respectively. Columns 2, 5, 8, and 11 show the cutoffs C estimated with the smoothed procedures for each group and depth, respectively. Columns 3, 6, 9, and 12 show the

DOI: 10.1002/env

Table 3. Outliers detected by the proposed procedure with the three considered depths. First part of the table shows the case of working days and second part shows the nonworking days

	Working days						Nonworking days					
	Trimming			Weighting			Trimming			Weighting		
	C	Depth	Out.	*C*	Depth	Out.	*C*	Depth	Out.	*C*	Depth	Out.
FMD	12.42	12.06	03/18	12.38	12.06	03/18	12.53	12.31	03/19	12.50	12.31	03/19
MD	0.97	0.68	03/18	1.04	0.68	03/18	0.97	0.87	03/19	0.93	0.87	03/19
		0.89	04/29		0.89	04/29		0.78	04/30		0.78	04/30
RPD	2.43	1.69	03/18	2.52	1.71	03/18	1.60	1.46	03/19	1.54	—	—
		2.24	04/29		1.81	04/29		1.44	04/30		—	—

value of the depths for the outliers detected by the procedure, which are shown in columns 4, 7, 10, and 13. With the FMD, the procedure, with both smoothed bootstrap estimates of the cutoff, detects one outlier in each group of curves in two consecutive days, 18 and 19 March. With the MD, the procedure detects, with both smoothed bootstrap estimates of the cutoff, two outliers in each group of curves in two consecutive days, 18 March and 29 April, and, 19 March and 30 April, respectively, for working and nonworking days. Finally, with the RPD, the procedure detects two outliers in each group of curves except for the case of nonworking days and the smoothed bootstrap based on weighting. In this latter case, if appears that the derivatives are not given much information about the outliers.

Once the outliers have been detected, we look for the causes for the abnormal values obtained by these curves. These causes may provide some information about the sources which produce abnormal large NO_x emissions. In this case, the answers are easy to find. The Friday, 18 March and Saturday, 19 March correspond to the beginning of the Eastern vacation in Spain in the year 2005. The Friday,

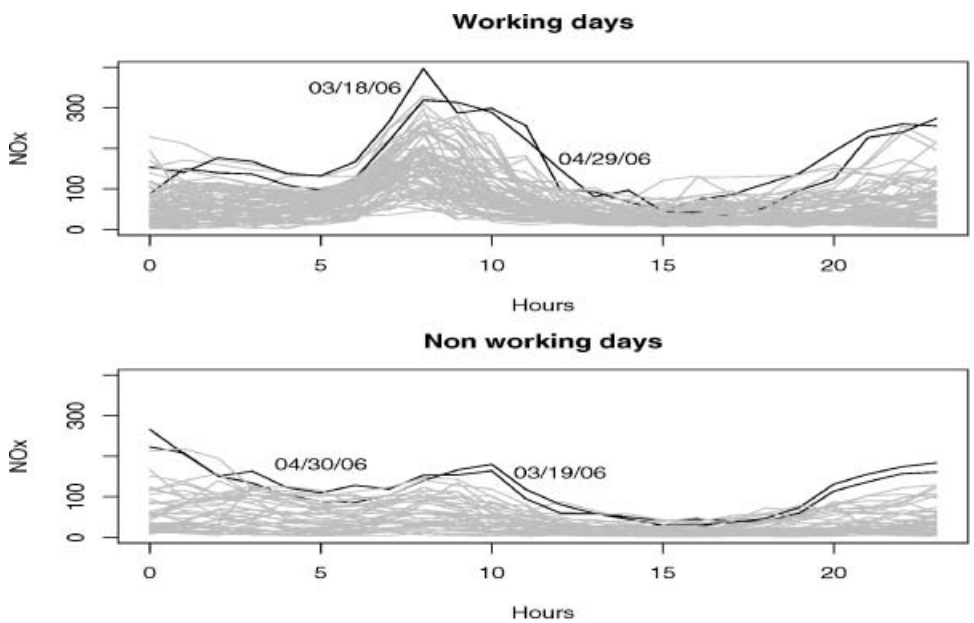

Figure 3. Outliers detected in the NO_x levels split into two groups: (up) working days; (down) nonworking days

DOI: 10.1002/env

29 April and Saturday, 30 April correspond to the beginning of a long weekend. Therefore, these outliers are related to small vacation periods which produces large traffic concentration in specific periods. We conclude that the abnormal observations detected are linked to a strong increase in traffic due to these vacation periods. Figure 3 shows the observed curves divided into two groups with the outliers detected by the procedure. The plot appears to confirm that the curves detected by the procedure as outliers show some abnormal behavior compared with the rest of curves during some part of the observation period.

Finally, in order to strengthen the convenience of using a functional method to detect outliers in functional datasets, we use the fast minimum covariance determinant (MCD) estimator of Rousseeuw and Van Driessen (1999) to obtain robust location and scatter estimates of both working and nonworking datasets. Then, robust Mahalanobis distances of the curves can be obtained using these robust estimates, which can be compared with the usual cutoff based on the χ^2 distribution in order to detect multivariate outliers. See Rousseeuw and Van Driessen (1999) for a detailed exposition of the method. Using this multivariate procedure, we detect 25 functional outliers for the working days dataset, which in view of the data, appears that it has no sense. On the other hand, the procedure cannot be applied to the nonworking days dataset, as the number of days in the data (39) is smaller than twice the number of observations (24).

ACKNOWLEDGEMENTS

The first and the third authors acknowledge financial support from grant MTM2005-00820. The second author acknowledges financial support by Xunta de Galicia under the Isidro Parga Pondal Program and the project PGIDIT06PXIB207009PR, and Ministerio de Educación y Ciencia grant SEJ2004-03303.

REFERENCES

Barnett V, Lewis T. 1994. *Outliers in Statistical Data* (3rd edn). Wiley: Chichester, England.

Boente G, Fraiman R. 2000. Kernel-based functional principal components. *Statistics and Probability Letters* **48**: 335–345.

Cai TT, Hall P. 2006. Prediction in functional linear regression. *Annals of Statistics* **34**: 2159–2179.

Cardot H, Ferraty F, Sarda P. 1999. Functional linear model. *Statistics and Probability Letters* **45**: 11–22.

Cuesta-Albertos JA, Fraiman R. 2007. Impartial trimmed k-means for functional data. *Computational Statistics and Data Analysis* **51**: 4864–4877. DOI:10.1016/j.csda.2006.07.011.

Cuevas A, Febrero M, Fraiman R. 2001. Cluster Analysis: a further approach based on density estimation. *Computational Statistics and Data Analysis* **36**: 441–456.

Cuevas A, Febrero M, Fraiman R. 2003. Linear functional regression: the case of fixed design and functional response. *Canadian Journal of Statistics* **30**: 285–300.

Cuevas A, Febrero M, Fraiman R. 2004. An anova test for functional data. *Computational Statistics and Data Analysis* **47**: 111–122.

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

Cuevas A, Febrero M, Fraiman R. 2007. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics* (in press). DOI:10.1007/s00180-007-0053-0.

Cuevas A. Fraiman R. 1997. A plugin approach to support estimation. *Annals of Statistics* **25**: 2300–2312.

Ferraty F, Vieu P. 2003. Curves discrimination: a nonparametric functional approach. *Computational Statistics and Data Analysis* **44**: 161–173.

Ferraty F, Vieu P. 2006. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer-Verlag: London.

Fraiman R, Muniz G. 2001. Trimmed means for functional data. *Test* **10**: 419–440.

Hall P, Hosseini-Nasab M. 2006. On properties of functional principal components analysis. *Journal of the Royal Statistical Society B* **68**: 109–126.

Liu R. 1990. On a notion of data depth based on random simplices. *Annals of Statistics* **18**: 405–414.

López-Pintado S, Romo J. 2006. On the concept of depth for functional data. *Working Paper 06-30*, Universidad Carlos III de Madrid, Madrid, Spain.

Oja H. 1983. Descriptive statistics for multivariate distributions. *Statistics and Probability Letters* **1**: 327–332.
Ramsay JO, Silverman BW. 2004. *Applied Functional Data Analysis*. Springer: New York.
Ramsay JO, Silverman BW. 2005. *Functional Data Analysis* (2nd edn). Springer: New York.
Rousseeuw PJ, Van Driessen K. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**: 212–223.
Silverman BW. 1996. Smoothed functional principal components analysis by choice of norm. *Annals of Statistics* **24**: 1–24.
Tukey JW. 1975. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, Vol. 2, James RD (ed.). Canadian Mathematical Society: Vancouver, 523–531.
Zuo Y. 2003. Projection based data depth functions and associated medians. *Annals of Statistics* **31**: 1460–1490.
Zuo Y, Serfling R. 2000. General notions of statistical depth function. *Annals of Statistics* **28**: 461–482.

DOI: 10.1002/env