A functional analysis of NOx levels: Location and scale estimation and outlier detection

Article *in* Computational Statistics · February 2007

DOI: 10.1007/s00180-007-0048-x · Source: RePEc

3 authors:

Manuel Febrero-Bande
University of Santiago de Compostela
78 PUBLICATIONS 1,969 CITATIONS
SEE PROFILE

Pedro Galeano
University Carlos III de Madrid
49 PUBLICATIONS 722 CITATIONS
SEE PROFILE

Wenceslao Gonzälez-Manteiga
University of Santiago de Compostela
235 PUBLICATIONS 3,523 CITATIONS
SEE PROFILE

UNIVERSIDADE DE
SANTIAGO DE COMPOSTELA
DEPARTAMENTO DE
ESTATÍSTICA E INVESTIGACIÓN OPERATIVA

A functional analysis of NOx levels: location and scale estimation and outlier detection

M. Febrero, P. Galeano, W. González Manteiga

Report 06-03

Reports in Statistics and Operations Research

A functional analysis of NOx levels: location and scale estimation and outlier detection

Manuel Febrero[1], Pedro Galeano[1] and Wenceslao González Manteiga[1]

[1]Departamento de Estadística e Investigación Operativa,
Universidad de Santiago de Compostela

Summary

This paper analyzes the NOx levels measured by a control station near a power plant by using techniques for functional data. First, we test for differences between the levels on working and non working days. Second, we obtain several location estimators and confidence sets of the center of the functional distribution. Third, we provide scale estimators and confidence sets of the dispersion of the functional distribution. Finally, a distance based procedure provides a criterion to determinate the presence of outlying observations, which allows to detect relevant NOx levels.

Keywords: Functional data analysis; Functional trimmed means; Functional trimmed standard deviation; NOx levels; Outliers.

1 Introduction

NOx (nitrogen oxides) is the term for a group of gases which mainly contents nitrogen and oxygen. Nitrogen oxides are, not only ones of the most important pollutants, but also ones of the precursors of ozone formation and contributors to global warming. Although NOx can be formed naturally, it is primarily caused by combustion processes in sources such us motor vehicles, electric utilities, industries and any other system that burn fuels. Nowadays, many governments have develop directives to put limit values for NOx emissions which mainly affect industries, airports and motor vehicles, among others. Therefore, it is necessary to develop procedures to study NOx emissions, for instance, in order to know if the NOx levels are different at different times of the week or if, occasionally, the levels are significatively large or small due to some abnormal effects.

The main purpose of this paper is to analyze the NOx levels measured by an environmental control station by means of techniques for functional data analysis, hereafter FDA. FDA is concerned with the analysis of functional random variables. We say that X is a functional random variable if it takes values in an infinite dimensional space. In the particular case of the NOx levels here analyzed, the observation space is a closed interval $[t_{\min}, t_{\max}]$ and the variable is observed at a discretized set of different times $t_1, \ldots, t_m$ belonging to $[t_{\min}, t_{\max}]$, providing an observation of the functional variable, $x(t_1), \ldots, x(t_m)$. Therefore, a dataset of n independent and identically distributed functional variables values, $x_1, \ldots, x_n$, observed at a grid of points $t_1, \ldots, t_m$ is given by,

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

Monographs on FDA can be found in Ramsay and Silverman (2004, 2005), which presents a large variety of methods and case studies and Ferraty and Vieu (2006), which presents a non parametric approach to analyze functional data. See also the references therein.

The dataset of NOx emissions that we have at hand have been taken by a control station in Barcelona, Spain, during the first semester of 2005 and can be downloaded from the webpage http://www.gencat.net. In particular the levels have been observed at every hour of every day of the observation period providing a long sample of measures that we split in functional samples of 24 hours observations. Thus, each curve represents the evolution of the levels in one day. Our analysis is composed of four aspects. First, we carry out an exploratory analysis of the data which allows us to characterize the behavior of the observed NOx levels as of a functional nature. After that, we analyze if there are differences between the levels on working and non working days by using an anova test for functional data. Second, we locate the center of the functional distribution of the NOx levels by means of location estimates and confidence sets. Third, we analyze the dispersion of the sample with two scale estimators: the functional standard deviation and the trimmed standard deviation, which, as far as we know, is firstly analyzed here in functional settings. Finally, it is important to identify days or periods in which the NOx levels are abnormally large or small, because these outlying observations may allow us to find out sources which produce large NOx emissions. Thus, we develop a distance based method for outlier detection in functional data which relies on a bootstrap procedure which allows to obtain percentiles of the distribution of functional distances of the curves with respect to a location estimator. If this distance for a curve relative to one scale estimator is large enough compared with the ones for the rest of curves, we assume that the curve is a functional outlier.

The rest of this paper is organized as follows. In section 2, we present the collected data and summarize their principal characteristics. In section 3, we analyze the presence of two groups of curves in the data. In section 4 we study location estimates and confidence sets for the NOx data, while in section 5 we obtain scale estimates, including the trimmed standard deviation for functional data. In section 6, we look for outliers in the NOx data by means of a distance based method. These outliers represent days in which the NOx levels are significatively large or small compared with the rest of the sample. Finally, in section 7, we conclude.

2 The NOx data

The data correspond to levels measured by a environmental control station in Poblenou, a neighborhood in Barcelona, which is around an industrial area in Besòs. The control station measures NOx levels in $\mu g/m^3$ every hour of every day. The dataset consists of the amount of 127 days of data, from February, 23th, to June, 26th in 2005. Figures 1 and 2 show boxplots of the data in terms of two factors: the hours (Fig. 1) and the day of the week (Fig. 2). The first graphic shows the boxplot of the data for the 24 hours of the day. The graphic gives us a first look of the behavior of the data. The NOx levels increase in the morning, attaining their largest values around 8:00am. Then the levels decrease until 14:00am approximately and increase again at the evening. As the control station is located at the city center, apparently there is a large influence of traffic on the measured NOx levels, as, on the other hand, it may be expected. The boxplot also suggest the presence of several outliers in most of the hours. The second graphic shows the boxplots of the data for two different subsamples: (1) working days, which are the weekdays, and (2) non working days, which are the Saturdays, Sundays and festive days. From the plot, we conclude that apparently may be differences on the levels of both groups. In both graphics, it is really meaningful the

Figure 1: Boxplot of the NOx data by hour of the day.

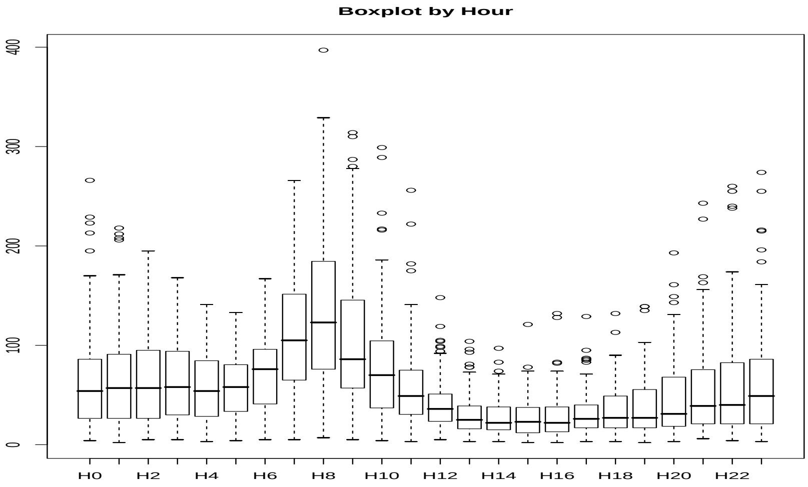


Figure 2: Boxplot of the NOx data by day of the week.

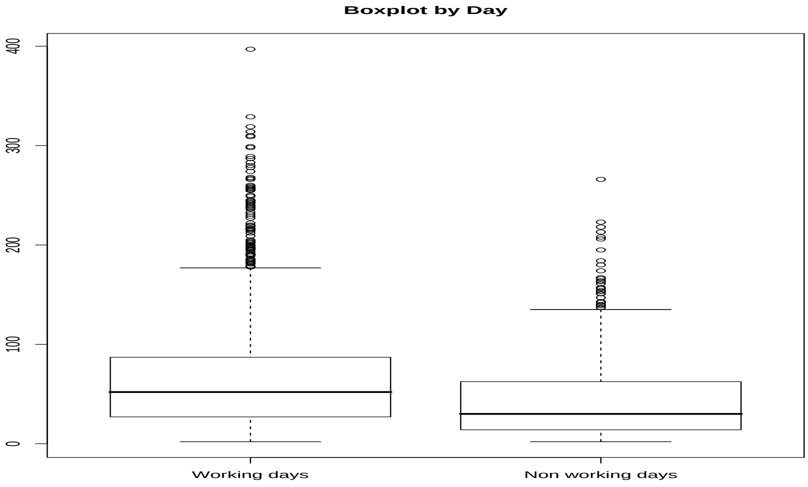
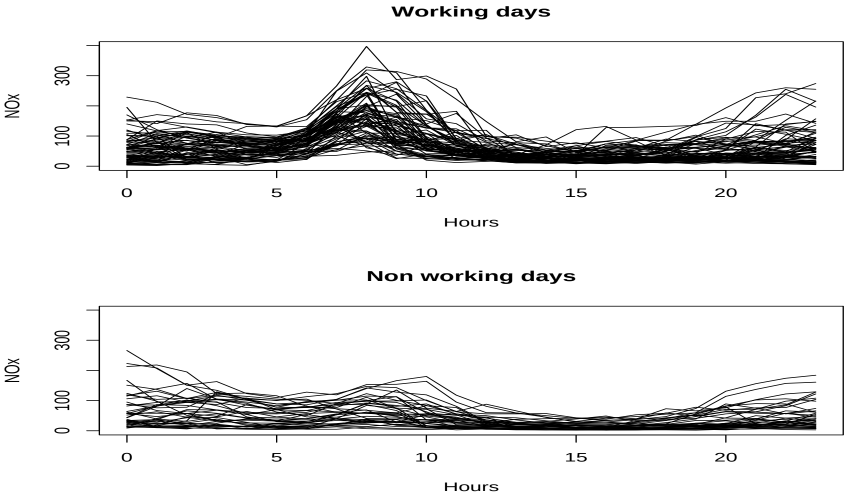


Figure 3: Sample curves of the NOx data: working days (up); non working days (down).



Working days

Non working days

presence of several extreme values, especially on working days. These both aspects, the existence of two groups and the presence of outliers, will be primarily the subjects of our ongoing analysis.

In order to make a functional analysis of the data, we assume that each curve is formed by the 24 observations of a day. Therefore, we have 127 curves corresponding to 127 days. Some of the measures are missing for several consecutive hours of some days, so that only 115 days are complete for the analysis. We decided to discard the days with incomplete periods. Figure 3 shows the 115 observed curves divided in two groups: working days (up) and non working days (down). As we can see, the shape of the observed curves are rather similar except for some of them that are different for the whole or some hours of the day. Note also that most of the largest values are attained for the working days, as expected from Figure 2.

3 Testing for the equallity of means

Once that we have introduced the data, the first step in our analysis is to determinate if there are significant differences between the levels depending on the day of the week. For that, we split the dataset in the two group of curves formed by the working and the non working days, and use the anova test for functional data proposed by Cuevas, Febrero and Fraiman (2004), for which we first need to introduce three important notions, which will be further used along the paper: sample mean, sample covariance and norms for functional data. Let $x_1(t), \ldots, x_n(t)$ be the $n = 115$ sample curves. The functional sample mean for the n curves is given by:

$$\widehat{\mu}_M(t) = \frac{1}{n} \sum_{i=1}^{n} x_i(t), \tag{1}$$

while the functional sample covariance is a matrix with size $m \times m$ and elements is given by:

$$\widehat{\Sigma}(t_j, t_k) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i(t_j) - \widehat{\mu}_M(t_j))(x_i(t_k) - \widehat{\mu}_M(t_k)),$$

for $j, k = 1, \ldots, m$. The sample functional mean is an estimator of the center of the functional distribution, while the sample functional covariance is an estimator of the scale and correlation structure of the functional distribution. Both estimators will be further analyzed for the NOx data in sections 4 and 5. Although several norms for functional data have been proposed, the most relevant ones are the L^p norms, where $p = 1, 2, \ldots, \infty$, which, for a curve $x_1(t)$, are given by:

$$\|x_1\|_p = \left(\int_{t_{\min}}^{t_{\max}} |x_1(t)|^p \, dt \right)^{\frac{1}{p}} \qquad \|x_1\|_\infty = \sup_{t \in (t_{\min}, t_{\max})} |x_1(t)|.$$

The distances between two curves $x_1(t)$ and $x_2(t)$ associated with these norms, $d_p(x_1, x_2)$, $p = 1, 2, \ldots, \infty$, are defined in the usual way as the norm of the functional difference, $x_1(t) - x_2(t)$. For more information and definitions of alternative distances see Ferraty and Vieu (2006).

The anova test proposed by Cuevas, Febrero and Fraiman (2004) is based on the statistic:

$$V = \sum_{g<h}^{G} n_g \|\widehat{\mu}_{M,g} - \widehat{\mu}_{M,h}\|_2, \tag{2}$$

where G is the number of groups, n_g is the number of curves in group g and $\widehat{\mu}_{M,g}$ denotes the functional sample mean for the curves in group g. We can conclude that the G groups are different if the statistic V is large enough. The asymptotic distribution of the statistic (2) can be approximated by means of a Monte Carlo procedure which, in the homoscedastic case, depends on the functional sample covariance matrix of the curves, and, in the heteroscedastic case, depends on the functional sample covariance matrices of the curves in each group $g = 1, \ldots, G$. We refer to Cuevas, Febrero and Fraiman (2004) for a more detailed exposition of the properties of the statistic (2) and its asymptotic distribution.

For the NOx data, we have $G = 2$ groups with $n_1 = 76$ and $n_2 = 39$ curves in each group, respectively. We apply the anova test assuming heteroscedasticity. The results indicated a strong evidence of the hypothesis that the two group of curves are different as the resulting p-value was 0. We conclude that there are significatively differences between the working and non working days. This conclusion may be expected because, as we have seen in Figures 1 and 2, the traffic appears to have a large influence on NOx levels. One may wonder if we can go further and look for differences between Saturdays and the subgroup of Sundays and festive days. Note that there are only 18 and 21 curves in the first and second subgroups, respectively, which is less than the 24 hours of the day. Thus, results obtained with these small groups may be unreliable, so we prefer to consider them as members of the same group.

Consequently, in what follows, we are going to do a parallel analysis of the NOx data, for the whole dataset and for both groups by separate. The rest of our analysis explores three different aspects: (1) location estimation; (2) scale estimation; and (3) outlier detection.

4 Location estimation: mean, trimmed mean, median and mode for functional data

Our second step in analyzing the NOx data is to provide location estimators and confidence sets of the central curve. Obviously, the first candidate to estimate the center of the distribution is the sample mean defined in (1), introduced in the previous section, but alternative location estimates have been proposed. As an attempt to obtain a robust location estimator of the center of the distribution, Fraiman and Muniz (2001) introduced the functional $\alpha-$trimmed mean which is defined as the mean of the most central $n - [\alpha n]$ curves, where α is such that $0 \leq \alpha \leq (n-1)/n$ and [ ] denotes the integer part. The notion of depth is used to define what the most central curve means. Depths for multivariate data points were introduced to measure the centrality of a multivariate observation within a given data cloud. For continuous one-dimensional random variables, the most popular depths are the halfspace depth, proposed by Tukey (1975), which, for a point x, drawn from a random variable with distribution function F, is given by:

$$HD(x) = \min\{F(x), 1 - F(x)\},$$

and the simplicial depth, proposed by Liu (1990), which is given by:

$$SD(x) = 2F(x)(1 - F(x)).$$

Also, Fraiman and Muniz (2001) considered a depth of the form:

$$FMD(x) = 1 - \left|\frac{1}{2} - F(x)\right|. \tag{3}$$

In practice, the distribution function F is substituted by the empirical distribution function of the observed sample. Thus, if D is a univariate depth defined on $\Re$, the univariate depth of the point $x_i(t)$ with respect to the sample points $x_1(t), \ldots, x_n(t)$, is given by $D(x_i(t))$, which allows to define the functional depths of the curves $x_1(t), \ldots, x_n(t)$ as follows:

$$FD(x_i(t)) = \int_{t_{min}}^{t_{max}} D(x_i(t))\,dt, \qquad i = 1, \ldots, n.$$

Therefore, each curve $x_i(t)$ is associated with its corresponding functional depth $FD(x_i(t))$, such that the deepest and the less deepest curves are the ones with attains the maximum and minimum values of the functional depths. If the curves are ranked according to decreasing values of their depths, we get the ordered curves $x_{(1)}(t), \ldots, x_{(n)}(t)$, such that $x_{(1)}(t)$ is the deepest curve and $x_{(n)}(t)$ is the less deepest one. The functional trimmed mean of $x_1(t), \ldots, x_n(t)$ for a given value α, is defined as:

$$\widehat{\mu}_{TM,\alpha}(t) = \frac{1}{n - [\alpha n]} \sum_{i=1}^{n-[\alpha n]} x_{(i)}(t).$$

Note that the trimmed means range from the functional mean to the median by considering α from 0 to $(n-1)/n$. Thus, by product we obtain that a trimmed mean in conjunction with a depth measure, provide different alternative ways to define a functional median. In what follows, we consider the functional median using the depth (3) which is denoted by $\widehat{\mu}_{MED}(t)$.

Finally, Cuevas, Febrero and Fraiman (2006) extended the concept of mode to the functional framework. Their idea was to select the curve most densely surrounded by the rest of curves of the dataset. The functional mode, $\widehat{\mu}_{MOD}(t)$ of a set of curves $x_1(t), \ldots, x_n(t)$, is then defined as:

$$\widehat{\mu}_{MOD}(t) = \arg\max \left\{ g_h(x_i(t)), i = 1, \ldots, n \right\},$$

such that,

$$g_h(x_i(t)) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{\| x_j(t) - x_i(t) \|}{h} \right),$$

where $\| \ \|$ is a norm in the functional space and $K : \Re^+ \to \Re^+$ is a kernel function. Note that only the median and the mode are curves belonging to the dataset, while the mean and trimmed means are just linear combinations of all the curves.

We define the confidence set of a curve $x_1(t)$ at the confidence level β as the set of curves $c(t)$ which have the same distribution that $x_1(t)$ and such that:

$$CS(x_1(t)) = \left\{ c(t) : d(x_1(t), c(t)) < D_\beta \right\},$$

where D_β is such that $\Pr\left(d\left(x_1\left(t\right), c\left(t\right)\right) < D_\beta\right) = \beta$, and d is a functional distance. A sample of curves belonging to the confidence set of a location estimator $\widehat{\mu}\left(t\right)$, $CS\left(\widehat{\mu}\left(t\right)\right)$, can be obtained using the smoothed bootstrap approach proposed by Cuevas, Febrero and Fraiman (2006) which works as follows. Let $\widehat{\mu}\left(t\right)$ be a location estimate based on the curves $x_1\left(t\right), \ldots, x_n\left(t\right)$ observed at times $t_1, \ldots, t_m$. First obtain B standard bootstrap samples from the curves that we denote by $x_i^b\left(t\right)$, for $i = 1, \ldots, n$ and $b = 1, \ldots, B$. Then, obtain smoothed bootstrap samples:

$$y_i^b\left(t\right) = x_i^b\left(t\right) + z_i^b\left(t\right),$$

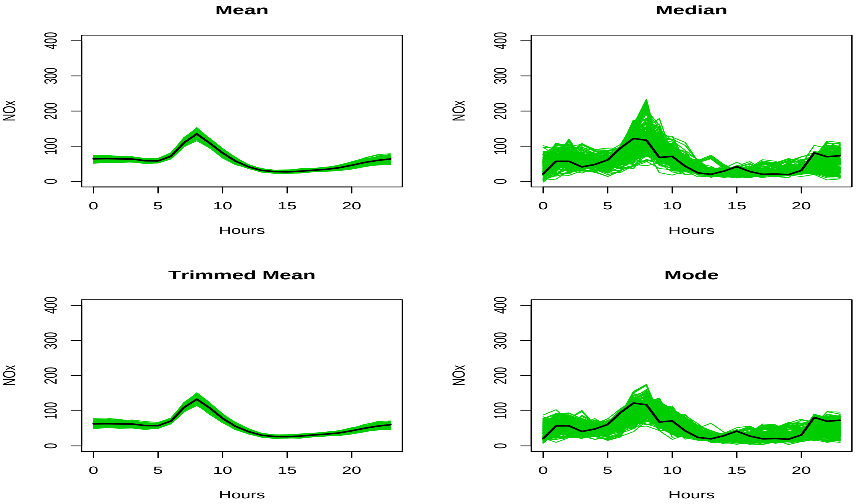
where $z_i^b\left(t\right)$ is such that $\left(z_i^b\left(t_1\right), \ldots, z_i^b\left(t_m\right)\right)$ is normally distributed with mean 0 and covariance matrix $\gamma\Sigma_x$, where Σ_x is the covariance matrix of $x\left(t_1\right), \ldots, x\left(t_m\right)$ and γ is a bootstrap smoothing parameter. Each smoothed bootstrap sample provides a location estimate, namely $\widehat{\mu}^b\left(t\right)$, such that a sample of curves belonging to $CS\left(\widehat{\mu}\left(t\right)\right)$ for confidence level β is defined by calculating the value D_B such that the $\left(100 \times \beta\right)\%$ of the smoothed boostrap replications $\widehat{\mu}^b\left(t\right)$ are within a distance from their average smaller than D_B. The sample of curves obtained with this procedure will be called a bootstrap confidence set.

Now, we obtain the location measures and their respective bootstrap confidence sets for the NOx data. Figure 4, 5 and 6 show the mean, median, trimmed mean and mode for the whole sample, the working days and the non working days, joint with their respective bootstrap confidence sets. In the three cases, we take $\alpha = 0.10$ to compute the trimmed mean, so that, the less 10% deepest curves are not taken into account for averaging. In order to compute the mode, as in Cuevas, Febrero and Fraiman (2006), we consider the Gaussian kernel:

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

and bandwidth $h = 0.2 \max\left\{\left\|x_j\left(t\right) - x_k\left(t\right)\right\|_2 : j, k = 1, \ldots, n\right\}$, where $\|\ \|_2$ is the L^2 norm. For the bootstrap confidence sets, we have taken $B = 200$ bootstrap samples, bootstrap smoothing parameter $\gamma = 0.05$, confidence level $\beta = 0.95$ and the L^2 distance. Some conclusions are as follows. First, the location estimates of the working and non working days are clearly different, as expected from the results in section 3. Second, note that the functional mean and trimmed mean are much less rough than the median and the mode. Obviously, this is due to the smoothing provided by the average of the curves in the computation of the means. Third, for the whole sample and the non working days, both the mean and trimmed mean and the median and mode, look very close to each other. This is not the case with the working days as the median and the mode strongly differ at the evenings, being the median closer to both means. Fourth, note also that the median and the mode for

Figure 4: Location estimates and bootstrap confidence sets for the whole sample.



the whole sample and the working days have a peak around 3:00pm, which is not observed for both means. This peak may be due to the traffic as this is the hour at which many people take their cars to come back home after work. Finally, the differences between the bootstrap confidence sets widths for the mean and trimmed mean with respect to the median and mode are quite large. This is also a consequence of the averaging made for the means, which considerably reduces the variability of these estimators in contrast with the median and mode.

5 Scale estimation: the standard deviation and trimmed standard deviation for functional data

As a third step in our analysis, we study the scale properties of the NOx data. The simplest candidate to estimate the dispersion of the curves is the sample standard deviation, which is defined by,

$$\widehat{\sigma}_{SD}(t) = \left(\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i(t) - \widehat{\mu}_M(t) \right)^2 \right)^{\frac{1}{2}}.$$

A more robust estimator of the dispersion of a univariate distribution is the

Figure 5: Location estimates and bootstrap confidence sets for the working days.

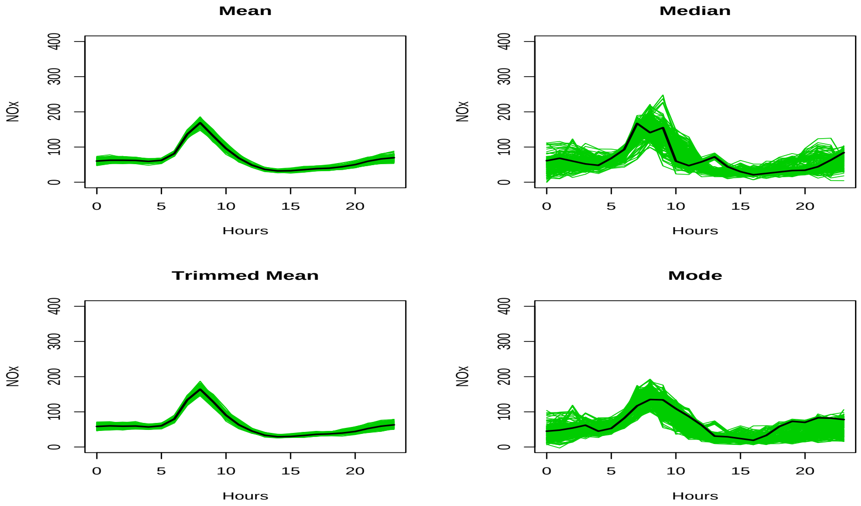
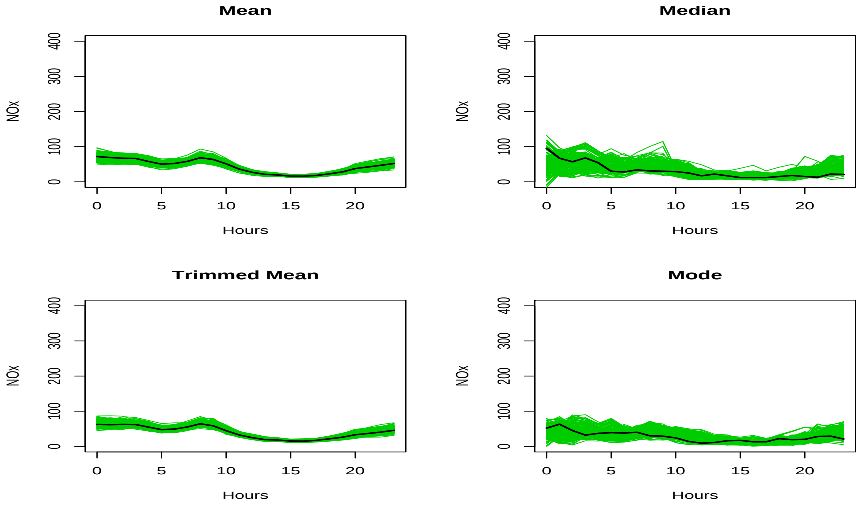


Figure 6: Location estimates and bootstrap confidence sets for the non working days.



trimmed standard deviation. The idea is similar to the trimmed mean: obtain the sample standard deviation of the deepest points. We generalize this estimator for functional settings by introducing the functional $\alpha-$trimmed standard deviation of $x_1(t), \ldots, x_n(t)$, which is defined as follows:

$$\widehat{\sigma}_{TSD,\alpha}(t) = \left(\frac{1}{n-[\alpha n]} \sum_{i=1}^{n-[\alpha n]} \left(x_{(i)}(t) - \widehat{\mu}_{TM,\alpha}(t) \right)^2 \right)^{\frac{1}{2}}.$$

As we are considering variation of the curves with respect to a trimmed mean, it is expected that $\widehat{\sigma}_{TSD,\alpha}(t)$ is less affected by extreme curves because the less deepest ones have no influence on the calculation of the trimmed standard deviation. As in the case of the trimmed mean, note that different trimmed standard deviations are defined by considering alternative depth measures. For both estimates, bootstrap confidence sets can be obtained with the bootstrap samples and the procedure described for location estimates.

Now, we obtain the scale measures and their respective bootstrap confidence sets for the NOx data. Figures 7 and 8 show the standard deviation and the trimmed standard deviation for the whole sample, the working days and the non working days, joint with their respective bootstrap confidence sets. In the three cases, we take $\alpha = 0.10$ to compute the trimmed standard deviation, so that, the less 10% deepest curves are not taken into account for averaging. The main conclusion from the plots is that the trimmed standard deviations in the three cases attain smaller values than the standard deviations. For instance, the reduction after trimming in the case of the working and non working days is as large as the 20%. Although it is expected that the trimmed deviation will give smaller values than the standard deviation, the amount of reduction getting here suggests that some curves which attain abnormal large values may be present in the sample. This conjecture is further analyzed in section 6. Finally, note that the bootstrap confidence sets widths are quite similar in both cases.

6 Outlier detection

The last step in our analysis is concerned with detection of outliers in the NOx data. Although the presence of outliers may have significative impact on FDA in many different ways, no outlier detection procedures have been proposed for functional data. Here, in order to look for outliers in the NOx data, we develop an algorithm based on distances and analyze its behavior for the NOx data. First of all, it is necessary to introduce some idea of what an outlier in functional settings is.

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 as-

Figure 7: Scale estimates and bootstrap confidence sets for the whole sample.

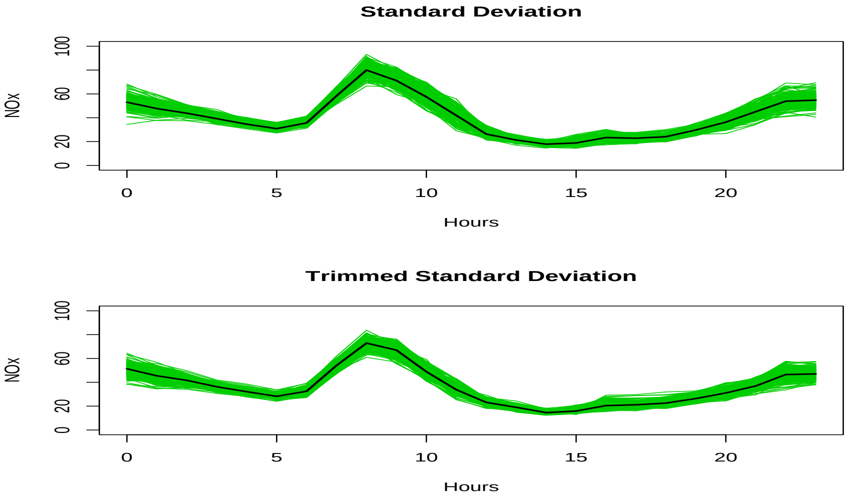
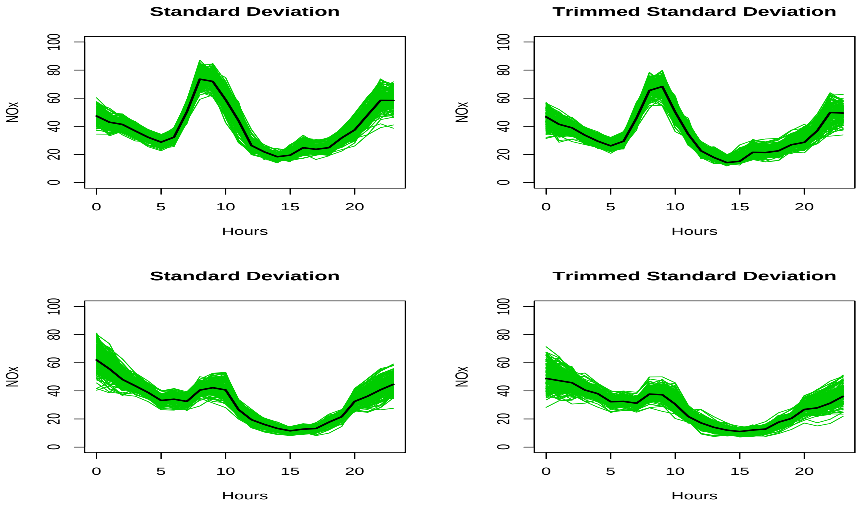


Figure 8: Scale estimates and bootstrap confidence sets for the working (up) and non working days (down).



sumed to be identically distributed. Note that this definition is wide enough to include curves which are different from the rest for all the observation period but also curves which are different from the rest only during some subinterval of the observation period.

It is well known that if $z_1, \dots, z_n$ is a sample drawn from the normal distribution, the likelihood ratio test statistic for testing that the observation z_i is an outlier is asymptotically the most powerful test, see, for instance, Barnett and Lewis (1994). This statistic is given by:

$$LRT\left(z_i\right) = \frac{z_i - \overline{z}}{\widehat{\sigma}}, \tag{4}$$

where $\overline{z}$ and $\widehat{\sigma}$ are the sample mean and standard deviation of the sample. In practice, the number and location of outliers are unknown a priori, so it is needed to check every observation for $i = 1, \dots, n$ and employ the statistic:

$$\lambda = \max_{1 \leq i \leq n} \left| LRT\left(z_i\right) \right|. \tag{5}$$

By comparing the test statistic (5) with some threshold, and an iterative procedure, one can determinate the presence of outliers. If the observations have been not drawn from the normal distribution, the likelihood ratio test statistic (4) can be seen as a quasi likelihood ratio test and still works well. Neither the sample mean nor the sample standard deviation are resistant to the presence of outliers, and this produces the effect known as "masking": a big outlier inflates the standard deviation, masking the presence of others. Thus, to avoid this effect, the mean and variance are replaced by some robust estimates such as the median or the trimmed mean, for the mean, and the median absolute deviation or the trimmed standard deviation, for the standard deviation.

In accordance with the reasoning for the univariate case, we proceed as follows. Let $O_\alpha\left(x_i\left(t\right)\right)$ be the statistic:

$$O_\alpha\left(x_i\left(t\right)\right) = \left\| \frac{x_i\left(t\right) - \widehat{\mu}_{TM,\alpha}\left(t\right)}{\widehat{\sigma}_{TM,\alpha}\left(t\right)} \right\|, \tag{6}$$

where $\| \; \|$ is a norm in the functional space ($\| \; \|_1$, $\| \; \|_2$ or $\| \; \|_\infty$), $\widehat{\mu}_{TM,\alpha}\left(t\right)$ is the $\alpha-$trimmed mean and $\widehat{\sigma}_{TSD,\alpha}\left(t\right)$ is the $\alpha-$trimmed standard deviation. Thus, $O_\alpha\left(x_i\left(t\right)\right)$ is the distance between $x_i\left(t\right)$ and $\widehat{\mu}_{TM,\alpha}\left(t\right)$ relative to $\widehat{\sigma}_{TSD,\alpha}\left(t\right)$. We look for functional outliers in the NOx data by using the statistic:

$$\Lambda = \max_{1 \leq i \leq n} O_\alpha\left(x_i\left(t\right)\right), \tag{7}$$

in conjunction with the following procedure:

Functional outlier detection procedure

1. Given the functional sample $x_1(t), \ldots, x_n(t)$, obtain the statistic (7).
2. Let $x_I(t)$ be the curve that attains the maximum value of the statistic (7). If $\Lambda = O_\alpha(x_I(t)) > C$, assume that $x_I(t)$ is an outlier, remove it from the sample, and repeat steps 1. and 2., until no more outliers are found.

The key point in the application of the algorithm is to obtain the threshold C. For that we propose the following bootstrap procedure, which make use of the smoothed bootstrap samples needed to obtain the confidence intervals of the location and scale estimators, and works as follows:

Bootstrap procedure for the threshold

1. Let $y_i^b(t)$, $i = 1, \ldots, n$ and $b = 1, \ldots, B$, be the B smoothed bootstrap samples. For each $b = 1, \ldots, B$, obtain:

$$\Lambda^b = \max_{1 \leq i \leq n-[\alpha n]} I_\alpha\left(y_{(i)}^b(t)\right),$$

   where $y_{(i)}^b(t)$, $i = 1, \ldots, n$, are the ordered smoothed bootstrap curves according with their depths, where α is the one taken to obtain $\widehat{\mu}_{TM,\alpha}(t)$ and $\widehat{\sigma}_{TSD,\alpha}(t)$ in the outlier detection procedure.

2. The maximum value of the sample $\Lambda^1, \ldots, \Lambda^B$ is the threshold C used in step 2. of the functional outlier detection procedure.

It is important to note that we compute the values $\Lambda^1, \ldots, \Lambda^B$ using only the $n - [\alpha n]$ most deepest smoothed bootstrap curves. This is done in order to avoid the presence of outliers in the bootstrap curves. But, if the dataset has no outliers, this choice may be not appropriate because the threshold C will be downward biased. Thus, we try to avoid the detection of false outliers by taking the threshold as the maximum of the set $\Lambda^1, \ldots, \Lambda^B$, which is expected to be large enough.

We apply the outlier detection procedure for the NOx data. Table 1 shows the outliers detected by the procedure with the three norms $\| \ \|_1$, $\| \ \|_2$ and $\| \ \|_\infty$. Rows 3 to 6, 7 to 9 and 10 to 11 show the outliers detected by the procedure for the whole sample, the working and non working days, respectively. Columns 2, 5 and 8 show the threshold obtained with the bootstrap procedure for each dataset and norm, respectively. Columns 3, 6 and 9 shows the values of the statistic Λ for the outliers detected by the procedure, which are shown in columns 4, 7 and 10. Note that the outliers detected with the three norms coincide for the working and non working days. This does not happen for

Table 1: Outliers detected by the procedure for the NOx data.

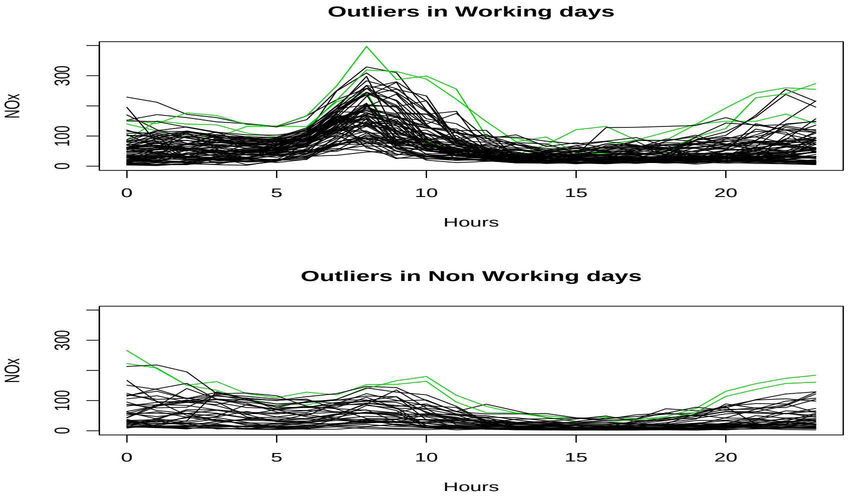
	$\| \|_1$			$\| \|_2$			$\| \|_\infty$		
Dataset	C	Λ	Outliers	C	Λ	Outliers	C	Λ	Outliers
Whole Sample	42.57	77.71	03/18	10.68	16.78	03/18	4.87	5.96	03/11
		61.20	04/29		14.42	04/29		5.94	03/18
		49.48	03/11		12.52	03/11		5.14	04/29
					10.87	05/02		5.06	05/02
Working Days	47.12	76.46	03/18	11.83	16.47	03/18	4.90	6.01	03/11
		60.38	04/29		14.26	04/29		5.61	03/18
		50.08	03/11		12.48	03/11		5.09	04/29
Non Working Days	55.84	61.63	04/30	11.23	13.76	04/30	3.98	4.43	04/30
		64.16	03/19		13.55	03/19		4.12	03/19

the outliers detected in the whole sample. This is not surprising as the whole sample is formed by two different groups of curves and the outliers have been detected not taken this fact into account. About the days in which the outliers has been detected, the Friday, 03/18 and Saturday, 03/19 correspond to the beginning of the Eastern vacation in Spain in the year 2005. The Friday, 04/29, Saturday, 04/30 and Monday, 05/02 correspond to a long weekend. Also the Friday, 03/11 is the beginning of a weekend. All these periods of time are related with vacation days, so that we conclude that the abnormal observations detected are linked to a strong increase in traffic due to small vacation periods. Figure 9 shows the observed curves divided in the two groups with the outliers detected by the procedure with the three norms. This plot confirms the results obtained by the proposed algorithm.

7 Conclusions

In this paper, we have analyzed a dataset of NOx emissions by using techniques for functional data analysis. First, we have found differences between the means of the groups formed of working and non working days by using an anova test. Second, several location estimates, including the mean, the median, the trimmed mean and the mode have been analyzed for the NOx emissions, joint with their respective confidence bands, which show differences between the estimates. Third, two scale estimates, including the standard deviation and the trimmed standard deviation have been obtained for the NOx emissions, joint with their respective confidence bands, suggesting the presence of outliers. Finally, we have found outliers in the NOx data by using an outlier detection procedure for functional data. All the outliers detected are linked to small vacation periods producing large traffic concentrations.

Figure 9: Outliers in the curves of the NOx data: working days (up) and non working days (down). Outliers are in green.



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 Ministerio de Educación y Ciencia grant SEJ2004-03303.

References

Barnett, V. and Lewis, T. (1994) Outliers in statistical data, 3rd Edition. John Wiley & Sons, Chichester.

Cuevas, A., Febrero, M. and Fraiman, R. (2004) An anova test for functional data. Computational Statistics and Data Analysis, 47, 111-122.

Cuevas, A., Febrero, M. and Fraiman, R. (2006) On the use of the bootstrap for estimating functions with functional data. Computational Statistics and Data Analysis, In press.

Ferraty, F. and Vieu, P. (2006) Nonparametric Functional Data Analysis: methods, theory, applications and implementations. Springer-Verlag, London.

Fraiman, R. and Muniz, G. (2001) Trimmed means for functional data. Test, 10, 419-440.

Liu, R. (1990) On a notion of data depth based on random simplices. Annals of Statistics, 18, 405-414.

Ramsay, J. O. and Silverman, B. W. (2004) Applied Functional Data Analysis. Springer, New York.

Ramsay, J. O. and Silverman, B. W. (2005) Functional Data Analysis, 2nd Edition. Springer, New York.

Tukey, J. W. (1975) Mathematics and the picturing of data. Proceedings of the International Congress of Mathematicians (R. D. James, Ed.), Vol. 2, pp. 523-531, Vancouver, 1975.

Reports in Statistics and Operations Research

2004

04-01 Goodness of fit test for linear regression models with missing response data. *González Manteiga, W., Pérez González, A.*
Canadian Journal of Statistics (to appear).

04-02 Boosting for Real and Functional Samples. An Application to an Environmental Problem. *B. M. Fernández de Castro and W. González Manteiga*.

04-03 Nonparametric classification of time series: Application to the bank share prices in Spanish stock market. *Juan M. Vilar, José A. Vilar and Sonia Pértega*.

04-04 Boosting and Neural Networks for Prediction of Heteroskedatic Time Series. *J. M. Matías, M. Febrero, W. González Manteiga and J. C. Reboredo.*

04-05 Partially Linear Regression Models with Farima-Garch Errors. An Application to the Forward Exchange Market. *G. Aneiros Pérez, W. González Manteiga and J. C. Reboredo Nogueira.*

04-06 A Flexible Method to Measure Synchrony in Neuronal Firing. *C. Faes, H. Geys, G. Molenberghs, M. Aerts, C. Cadarso-Suárez, C. Acuña and M. Cano.*

04-07 Testing for factor-by-curve interactions in generalized additive models: an application to neuronal activity in the prefrontal cortex during a discrimination task. *J. Roca-Pardiñas, C. Cadarso-Suárez, V. Nacher and C. Acuña.*

04-08 Bootstrap Estimation of the Mean Squared Error of an EBLUP in Mixed Linear Models for Small Areas. *W. González Manteiga, M. J. Lombardía, I. Molina, D. Morales and L. Santamaría.*

04-09 Set estimation under convexity type assumptions. *A. Rodríguez Casal.*

2005

05-01 SiZer Map for Evaluating a Bootstrap Local Bandwidth Selector in Nonparametric Additive Models. *M. D. Martínez-Miranda, R. Raya-Miranda, W. González-Manteiga and A. González-Carmona.*

05-02 The Role of Commitment in Repeated Games. *I. García Jurado, Julio González Díaz*.

05-03 Project Games. *A. Estévez Fernández, P. Borm, H. Hamers*

05-04 Semiparametric Inference in Generalized Mixed Effects Models. *M. J. Lombardía, S. Sperlich*

2006

06-01 A unifying model for contests: effort-prize games

06-02 The Harsanyi paradox and the "right to talk" in bargaining among coalitions

Previous issues (2001 – 2003):
http://eio.usc.es/pub/reports.html