UNIVERSIDAD CARLOS III DE MADRID

working papers

UC3M Working Papers
Statistics and Econometrics
17-12
ISSN 2387-0303
Mayo 2017

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-48

PREDICTION BANDS FOR FUNCTIONAL DATA BASED ON DEPTH MEASURES

Antonio Elías Fernández[a] , Raúl José Jiménez Recaredo[b]

Abstract

We propose a new methodology for predicting a partially observed curve from a functional data sample. The novelty of our approach relies on the selection of sample curves which form tight bands that preserve the shape of the curve to predict, making this a deep datum. The involved subsampling problem is dealt by algorithms specially designed to be used in conjunction with two different tools for computing central regions for functional data. From this merge, we obtain prediction bands for the unobserved part of the curve in question. We test our algorithms by forecasting the Spanish electricity demand and imputing missing daily temperatures. The results are consistent with our simulation that show that we can predict at the far horizon.

Keywords:* *depth measures, central regions, electricity demand, daily temperatures.

[a] Phd Student, Department of Statistics, Universidad Carlos III de Madrid.

[b] Department of Statistics, Universidad Carlos III de Madrid.

Acknowledgements: A. Elías is supported by the Spanish Ministerio de Educación, Cultura y Deporte under grant FPU15/00625. Both authors are partially supported by the Spanish Ministerio de Economía y Competitividad under grant ECO2015-66593-P.

Prediction bands for functional data based on depth measures

Antonio Elías[*] and Raúl Jiménez[†]
Department of Statistics, Universidad Carlos III de Madrid

May 24, 2017

Abstract

We propose a new methodology for predicting a partially observed curve from a functional data sample. The novelty of our approach relies on the selection of sample curves which form tight bands that preserve the shape of the curve to predict, making this a deep datum. The involved subsampling problem is dealt by algorithms specially designed to be used in conjunction with two different tools for computing central regions for functional data. From this merge we obtain prediction bands for the unobserved part of the curve in question. We test our algorithms by forecasting the Spanish electricity demand and imputing missing daily temperatures. The results are consistent with our simulation that show that we are able to predict at the far horizon.

Keywords: depth measures, central regions, electricity demand, daily temperatures.

[*] Supported by the Spanish Ministerio de Educación, Cultura y Deporte under grant FPU15/00625.

[†] Partially supported by the Spanish Ministerio de Economía y Competitividad under grant ECO2015-66593-P.

1 Introduction

The concept of depth for functional data, first discussed by Fraiman and Muniz (2001), has received a great deal of attention in recent years (Cuesta-Albertos and Nieto-Reyes, 2008; López-Pintado and Romo, 2009, 2011; Narisetty and Nair, 2015; Nieto-Reyes and Battey, 2016). Several functional depth measures have been considered for different application, including classification (Cuevas et al., 2007), outlier detection (Febrero et al., 2008; Arribas-Gil and Romo, 2014), measuring dispersion of curves and rank tests (López-Pintado et al., 2010). Functional versions of boxplots and other graphical tools based on different depths have been also proposed for visualizing curves with the aim of discovering features from a sample that might not be apparent by using other methods (Hyndman and Shang, 2010; Sun and Genton, 2011). The key instrument of these methods is the band delimited by the deepest curves, also termed central region. Central regions have been successfully used for shape and magnitude outliers detection. However, as far as we know, they have not been used for predicting unobserved part of curves, an important issue in the literature related to missing data (James et al., 2000; James and Hastie, 2001; Yao et al., 2005; Delaigle and Hall, 2013; Chiou et al., 2014) and forecasting (Antoniadis et al., 2006; Aneiros-Pérez and Vieu, 2008; Aneiros-Pérez et al., 2011; Shang and Hyndman, 2011; Shang, 2017).

To illustrate the power of central regions based on depths for forecasting, consider the average monthly sea temperatures between 1950 and 2015 measured by moored buoys in the "Niño region". The data was taken from `http://www.cpc.ncep.noaa.gov/`and have been already used by Hyndman and Shang (2010) and Sun and Genton (2011) to illustrate their methodologies. We will divide years in two halves and reserve the last six months for testing the prediction. Thus, considering data only from the first six months, we compute the deepest curve and the central regions C_α described by Sun and Genton (2011) for $\alpha = 0.1, 0.3$ and 0.5. In the next section we will detail which is this deepest curve and what is C_α; for now it is enough to think that the former is "in the middle" of the curves at most of the first half of the year and C_α is a band surrounding the deepest where $\alpha \times 100\%$ of the sample curves are enclosed during the first six months. The three central regions are differentiated by grey grades (black, dark and light grey) on the left side of Figure 1, where we also represent the temperatures of 1990 by a solid red line (corresponding to the

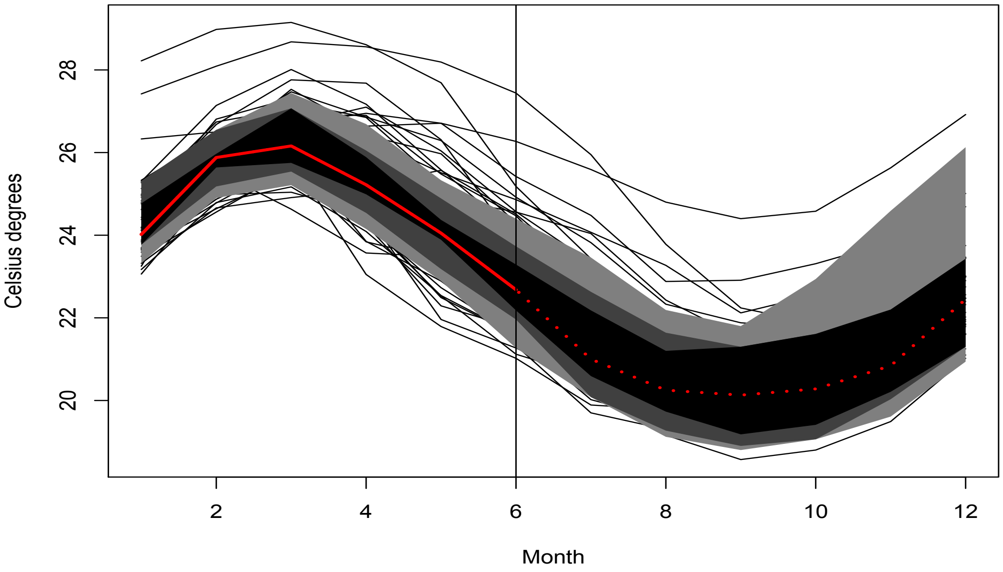

Figure 1: Curves of average monthly sea temperatures (1950–2015). Left-half side: the deepest curve on the first six months (corresponding to 1990, solid red line) and central regions C_α ($\alpha = 0.1, 0.3$ and 0.5, in black and two grey grades) computed according to Sun and Genton (2011) but considering data only from the first six months. Right-half side: extended central regions according to equation (1) and temperatures on the last six months of 1990 (dashed red line).

deepest curve on the first six months). Let $\mathcal{J}_\alpha$ be the set of curves that are completely contained in C_α on the first six months. Then, we extend the regions C_α to the last six months in the straightforward way by

$$\overline{C}_\alpha = \{(t, y(t)) : 6 < t \leq 12 \text{ and } \min_{y \in \mathcal{J}_\alpha} y(t) \leq y(t) \leq \max_{y \in \mathcal{J}_\alpha} y(t)\}. \tag{1}$$

As one can see in the Figure 1, the extended regions $\overline{C}_\alpha$ (identified by the same grey grades used for the corresponding C_α) envelope the dashed line providing prediction bands for the temperatures of the last six month of 1990. Smaller the value of α, tighter $\overline{C}_\alpha$. In addition, notice how $\overline{C}_\alpha$ preserves the shape of the dashed red line. In this article, we are interested in constructing similar prediction bands for the unobserved part of a curve, not necessarily the deepest. The main problem of this task consists on how to select sample curves surrounding the curve to extend. We propose two algorithms to address this problem which depend on the way of computing central regions, according to Hyndman and Shang (2010) and Sun and Genton (2011). James et al. (2000), James and Hastie (2001) and Yao et al. (2005) consider sparse functional data sets and deal with the problem of curve reconstruction based on parametric modeling. On other hand and in the context of functional data classification, Delaigle and Hall (2013) provide a nonparametric approach for extending curves by vertical translation of other pieces of functions from the sample. Rather than to extend an unobserved fragment by a piece of curve, Shang and Hyndman (2011) and Shang (2017) propose nonparametric methods for forecasting when the data come from a functional time series. In a more general framework, we introduce a new approach for providing prediction bands which preserve the shape of the function to predict.

This paper is organized as follows. In Section 2 the concept of central regions is reviewed. Section 3 presents two algorithms for subsampling curves surrounding a function and making the latter a deep datum. The prediction problem is considered in Section 4 and in Section 5 we apply the methodology proposed to two real data sets, the Spanish daily temperatures and electricity demand. Finally, we present some conclusions of our study in Section 6.

2 Functional Central Regions

Consider a sample of random functions $\mathcal{Y} = \{y_1, ..., y_n\}$ observed on a common compact interval I. For each $y \in \mathcal{Y}$, denote by $D(y, \mathcal{Y})$ the sample depth of y relative to $\mathcal{Y}$ according to some depth measure. We refer the reader to the recent paper of Nieto-Reyes and Battey (2016) for a thorough discussion on the definition of depth for functional data. Consider now the ordered sample $\{y_{(1)}, ..., y_{(n)}\}$ from larger to lower depth. This is $D(y_{(1)}, \mathcal{Y}) \geq D(y_{(2)}, \mathcal{Y}) \geq \dots D(y_{(n)}, \mathcal{Y})$. Then, the band delimited by the α proportion of deepest sample functions

$$C_\alpha = \{(t, y(t)) : t \in I, \min_{r=1,\dots,[\alpha n]} y_{(r)}(t) \leq y(t) \leq \max_{r=1,\dots,[\alpha n]} y_{(r)}(t)\}, \tag{2}$$

where $[\alpha n]$ is the integer part of αn, is referred to as the α sample central region (Sun and Genton, 2011). This concept is an extension of the definition of multivariate central region introduced by Liu et al. (1999).

In this paper, we will consider two ways of ordering functional data for computing central regions. By one hand, Hyndman and Shang (2010) first compute the cloud of points corresponding to the first two robust principal component scores of the sample curves. These scores are computed by applying the Croux and Ruiz-Gazen (2005) algorithm. Then, they assign to each curve the celebrated Tukey's halfspace depth of its score relative to the cloud of points. The Tukey's depth of a point w relative to a cloud of points $\mathcal{W}$, denoted here by $TD(w, \mathcal{W})$, is defined as the smallest number of points of the cloud contained in a closed half-plane containing w on its boundary (Tukey, 1975). By the other hand, Sun and Genton (2011) consider the modified band depth of López-Pintado and Romo (2009) with bands formed by two curves. Namely,

$$MBD(y, \mathcal{Y}) = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \frac{\lambda\left(\{t \in I : \min(y_i(t), y_j(t)) \leq y(t) \leq \max(y_i(t), y_j(t))\}\right)}{\lambda(I)},$$

λ being the Lebesgue measure.

Both ordering ways, based on TD and MBD, have demonstrated to provide central regions that envelope the deepest curve (with curves from below and above) and exhibit its shape. They have been also successfully used for outlier detection, specially to identify sample curves that lie outside the range of the vast majority of the data (magnitude outliers).

As an alternative of the central region based on the Tukey's depth, Hyndman and Shang (2010) also consider the so called functional *high density regions* (HDR). For this, they assign to each curve its bivariate kernel density estimate (Terrell and Scott, 1992) calculated from the first two robust principal component scores. Then, the functional HDR is defined as the band delimited by the sample curves whose corresponding scores are inside the bivariate region with coverage probability $1 - \alpha$, where all scores within the region have a higher density estimate than any of the points outside the region. The authors propose these regions having in mind curves that may be within the range of the rest of the data but have a very different shape from other curves (shape outliers). We remark that, if the bivariate kernel density involved is multimodal, the HDR may consist in an union of separate bands, as Hyndman and Shang (2010) have shown. Therefore, we discard HDR in this paper, we are interested in just one band that envelops curves surrounding the deepest, as the central regions discussed above.

3 Curves selection

Additionally to the n sample curves in $\mathcal{Y}$, consider a *focal curve* $y_0 \notin \mathcal{Y}$, the function that we are interested in predicting. We assume that y_0 is observed on $\mathcal{I}$, where the n sample curves are also observed. In line with the ideas sketched in the introduction, we are interested in subsamples $\mathcal{J} \subset \mathcal{Y}$ with tight central regions that envelope y_0. Of the 2^n possible subsamples, many of them may have tight central regions which scarcely cover the focal curve, even if they are nearby. Conversely, many subsamples may completely cover the focal with wide central regions, with boundaries faraway from y_0. These regions may not provide proper information about the features of the focal curve. To illustrate the above, consider the sample of random harmonic signals shown in Figure 2 (black lines) and the three subsamples (blue lines) of each panel. The central region delimited by the two deepest curves of each subsample is painted in gray and the focal curve in red. The band of the left panel completely cover the focal curve, but it is the widest possible sample band. The band of central panel is tight and delimited by the two nearest curves in L^2 to the focal, however the focal curve is always out of this band. In fact, curves faraway from above and below usually envelope the focal but provide wide bands. Also, the nearest curves provide tight bands but they may not contain important parts of focal. Our goal is to provide tight central regions like the shown in right panel, where the focal is completely enveloped. We will address the problem by subsampling curves *from* y_0 *to outwards*, making y_0 a deep datum. According to the widely-accepted statistical wisdom which suggest large samples to decrease sampling errors, we are also interested in such large subsamples as possible. For that, we introduce two heuristics inspired by the two methods for computing central regions that we discussed in the previous section.

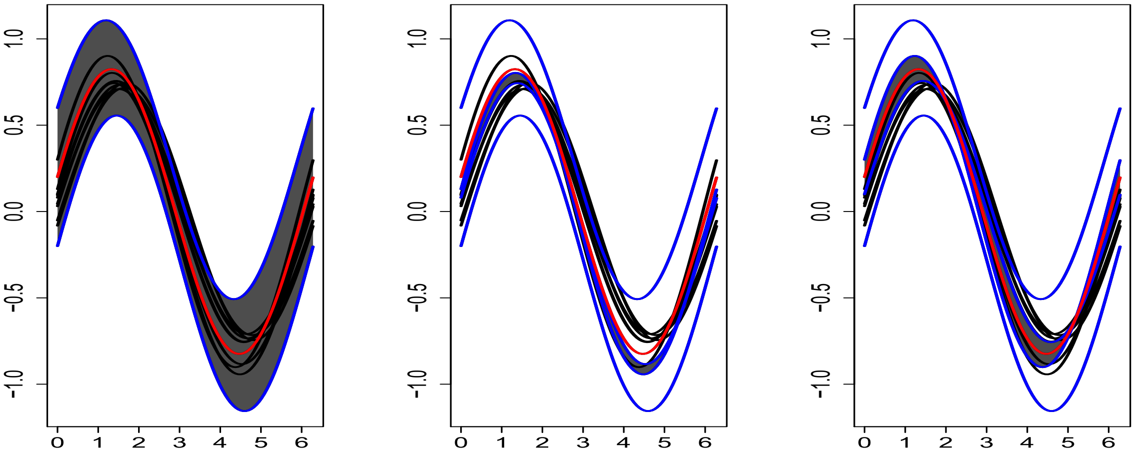

Figure 2: Three scenarios of central regions (dark gray) from subsamples (blue lines) of harmonic curves. Left panel: the region delimited by the farthest curves to the focal (in red). Central panel: the region delimited by the two nearest curves to the focal. Right panel: the tightest region that envelopes the focal.

3.1 Selection based on principal-component neighbourhoods

Following Hyndman and Shang (2010), we begin by mapping the curves $\{y_0, y_1, ..., y_n\}$ to their first two robust principal component scores denoted by $\mathcal{W}_0 = \{w_0, w_1, ..., w_n\}$. Since the main features of the curves should be captured by the scores (Jones and Rice, 1992), clusters of scores centered on w_0 should correspond to curves surrounding y_0. In order to construct these clusters, we consider Delaunay triangulations, probably the most widely used tool for solving proximity problems in stochastic geometry (Okabe et al., 2008). Henceforward, $\mathcal{D}(\mathcal{W})$ denotes the Delaunay triangulation on a set of scores $\mathcal{W}$.

Consider the neighbours of w_0 in $\mathcal{D}(\mathcal{W}_0)$. These neighbours are the scores directly connected to w_0 by an edge of $\mathcal{D}(\mathcal{W}_0)$. If w_0 is inside the polygon of minimum perimeter that joins its neighbours (the loop through these scores), we gather them in a set $\mathcal{N}_0$ for surrounding w_0. Next, we consider the triangulation on $\mathcal{W}_0 \setminus \mathcal{N}_0$. Let $\mathcal{N}$ be the neighbours of w_0 . If the loop through $\mathcal{N}$ contains w_0 and the percentile in depth of w_0 relative to $\mathcal{N}_0 \cup \mathcal{N}$ is greater or equal to the percentile relative to $\mathcal{N}_0$ then we add $\mathcal{N}$ to $\mathcal{N}_0$. Otherwise, $\mathcal{N}$ is discarded from $\mathcal{W}_0$. Here we use the Tukey's depth. We repeat the process until there are not loops around w_0. The final subsample is obtained by mapping the scores $\mathcal{N}_0$ to the set of sample functions $\mathcal{Y}$. This algorithm is describe below in pseudocode (Algorithm 1)

Algorithm 1 Input: $\mathcal{Y}, \mathcal{W}_0$/Output: $\mathcal{J}$

```
Initialize $\mathcal{N}_0 = \emptyset$
while $\mathcal{D}(\mathcal{W}_0 \setminus \mathcal{N}_0) \neq \emptyset$ do
    Determine the neighbours $\mathcal{N}$ of $w_0$ in $\mathcal{D}(\mathcal{W}_0 \setminus \mathcal{N}_0)$
    Compute the loop $\mathcal{L}$ through the points of $\mathcal{N}$
    Let $p_0$ be the percentile of $TD\,(w_0, \mathcal{N}_0 \cup \{w_0\})$
    Let $p_1$ be the percetile of $TD\,(w_0, \mathcal{N}_0 \cup \{w_0\} \cup \mathcal{N})$
    if $w_0$ is inside $\mathcal{L}$ and $p_1 \geq p_0$ then
        $\mathcal{N}_0 = \mathcal{N}_0 \cup \mathcal{N}$
    else
        $\mathcal{W}_0 = \mathcal{W}_0 \setminus \mathcal{N}$
    end if
end while
Compute $\mathcal{J}$ by mapping $\mathcal{N}_0$ to $\mathcal{Y}$
```

For illustration purposes, we consider again the sea surface temperatures discussed in Section 1 and we choose the curve corresponding to 1999 as focal. Figure 3 shows outputs of the first, second and last (third) iteration of Algorithm 1. Left panels display Delaunay triangulations (in grey), w_0 (in red), its neighbours and the corresponding loops (in black). Central panels show the expansion of $\mathcal{N}_0$ (in black) by iteration and right panels the selected curves (in black) for surrounding the focal (in red). Note how the corresponding scores of $\mathcal{J}$ increases at each iteration around w_0 by adding the vertices of polygons that contain w_0 (black points in left panel of Figure 3). Hence, the resulting cluster of scores (black points of bottom-center panels) form a cloud of points almost centered on w_0.

3.2 Selection based on band depth

Unlike the selection based on proximity of principal components scores, now we consider proximity of curves directly in the functional space. Remember that we look for subsample

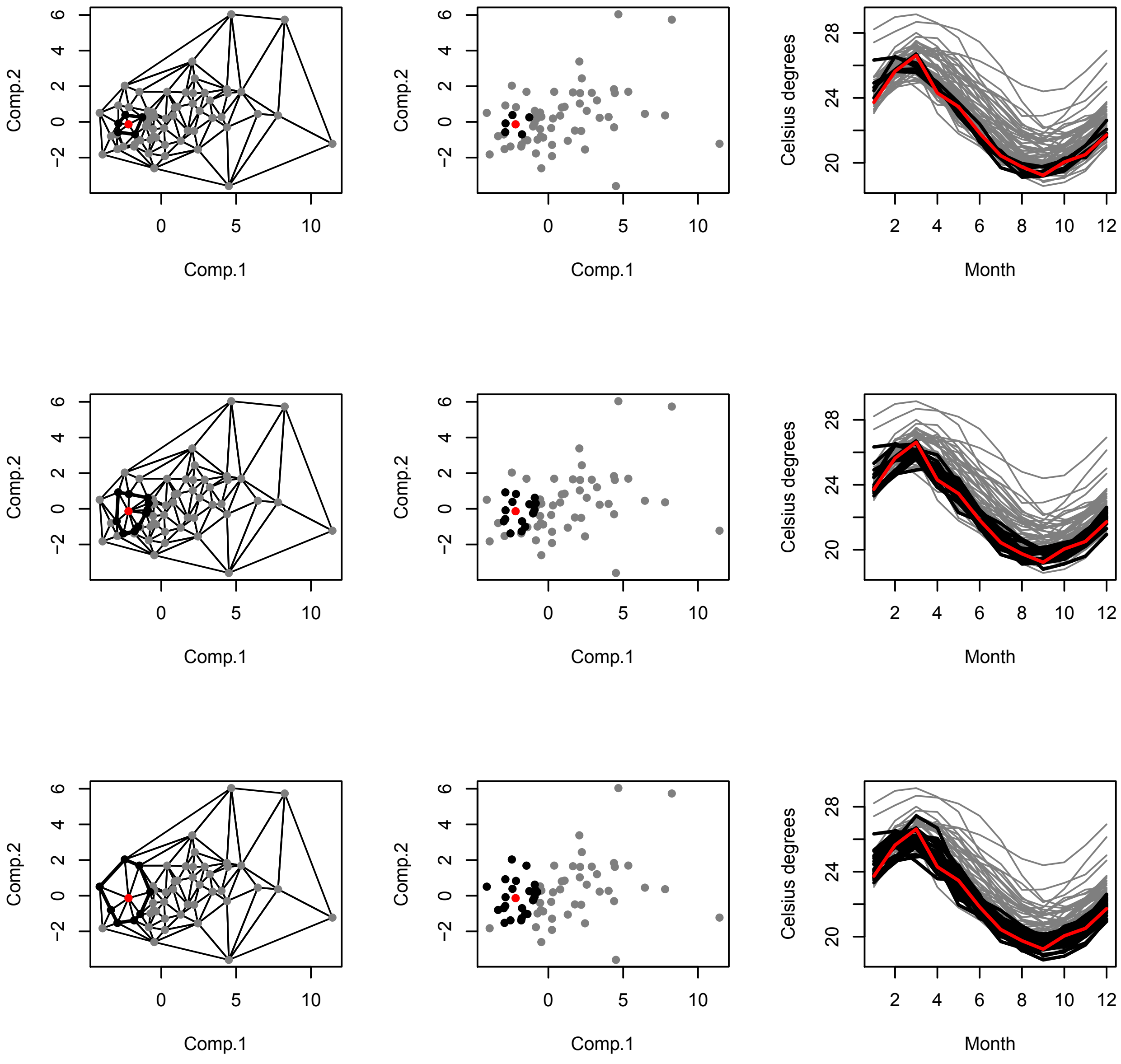

Figure 3: First, second and last (fourth) iteration of Algorithm 1 applied to the average monthly sea temperatures. Left panels: Delaunay triangulations (in grey), w_0 (in red), the triangles incident towards w_0 and its vertexes (in black). Central panels: $\mathcal{N}_0$ (in black). Right panels: selected curves (in black) for surrounding the focal (in red), corresponding to 1999.

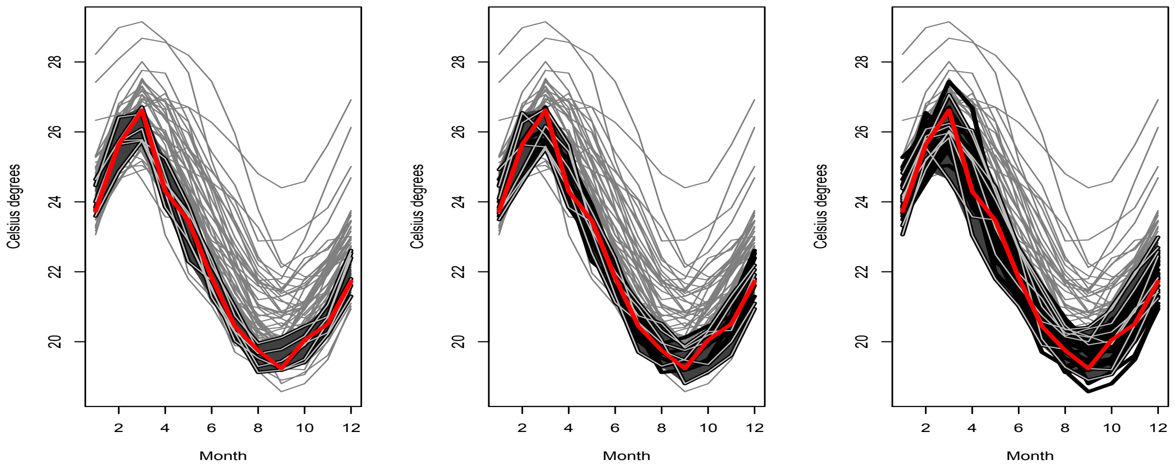

Figure 4: First, second and last (fourth) iteration of Algorithm 2 from average monthly sea temperatures. In dark gray the band delimited by the curves chosen at each iteration. In black, the curves collected previously. In red, the focal curve.

with central regions formed by nearby curves (here we use L^2-norm) which wrap the focal as much as possible. For doing so, first we identify the set $I_n \subset I$ where y_0 is covered by the sample. This is

$$I_n = \{t \in I : \min_{y \in \mathcal{Y}} y(t) \leq y_0(t) \leq \max_{y \in \mathcal{Y}} y(t)\}.$$

Then, we find the nearest sample curves to y_0 which envelope it on I_n. We gather these curves in a set $\mathcal{J}$. Next, we find the nearest curves in $\mathcal{Y} \setminus \mathcal{J}$ which envelope the focal as much as possible and collect them in $\mathcal{N}$. If the percentile in depth of y_0 relative to $\mathcal{J} \cup \mathcal{N} \cup \{y_0\}$ is greater or equal to the percentile relative to $\mathcal{J} \cup \{y_0\}$ then we add $\mathcal{N}$ to $\mathcal{J}$, otherwise we remove $\mathcal{N}$ from $\mathcal{Y}$ and repeat the process. In this procedure the depth used is MBD. This algorithm is detailed below and Figure 4 shows its first, second and last iterations from the sea surface temperatures, given 1999 as focal.

In order to compare the algorithms, we will vary the focal curve among the 66 years (1950–2015) and run both procedures. Denote by $\mathcal{J}(y)$ the subsample obtained for the focal curve y. Note that $\mathcal{J}(y)$ may be empty with Algorithm 1 if the score related to the principal components of y is out of any sample polygon. In contrast, Algorithm 2 provides bands excepting when y is always above or below the rest of the curves. We report the

Algorithm 2 Input: $\mathcal{Y}, y_0$/ Output: $\mathcal{J}$

```
Initialize $\mathcal{J} = \emptyset$
while  size of $\mathcal{Y} \setminus \mathcal{J} \geq 2$  do
    Let $y_{(k)}$ be the $k$th-nearest curve to $y_0$ from $\mathcal{Y} \setminus \mathcal{J}$.
    Let $\mathcal{N} = y_{(1)}$ and $m = 0$
    for $k \geq 2$  do
        $\lambda_k = \lambda\left(\{t \in I : \min_{y \in \mathcal{N} \cup y_{(k)}} y(t) \leq y_0(t) \leq \max_{y \in \mathcal{N} \cup y_{(k)}} y(t)\}\right)$
        if $\lambda_k > m$ then
            $\mathcal{N} = \mathcal{N} \cup \{y_{(k)}\}$
            $m = \lambda_k$
        end if
    end for
    Let $p_0$ be the percentile of $MBD\left(y_0, \mathcal{J} \cup \{y_0\}\right)$
    Let $p_1$ be the percetile of $MBD\left(y_0, \mathcal{J} \cup \{y_0\} \cup \mathcal{N}\right)$
    if  $m > 0$ and $p_1 \geq p_0$ then
        $\mathcal{J} = \mathcal{J} \cup \mathcal{N}_1$
    else
        $\mathcal{Y} = \mathcal{Y} \setminus \mathcal{N}_1$
    end if
end while
```

	Algorithm 1	**Algorithm 2**
outlying curves	8 out of 66	0 out of 66
subsample size	20.534	18.276
depth percentile	0.997	0.927

Table 1: Averages obtained by varying focal curve on the data set of sea temperatures.

number of such cases as *number of outlying curves*. For all the other cases, we compute the following performance statistics:

- *Subsample size*, number of curves in $\mathcal{J}(y)$.
- *Depth percentile*, the percentile of $D(y, \mathcal{J}^+(y))$ in $\{D(z, \mathcal{J}^+(y)), z \in \mathcal{J}^+(y)\}$, being $\mathcal{J}^+(y) = \mathcal{J}(y) \cup \{y\}$.

Consider also the k-deepest curves of $\mathcal{J}^+(y)$ and denote by $\mathcal{J}_k(y)$ these curves, excluding the focal if this is one of the k deepest. We also compute:

- *Standardized mean width*

  $$W_k(y) = \sum_i \left(\max_{z \in \mathcal{J}_k(y)} z(t_i) - \min_{z \in \mathcal{J}_k(y)} z(t_i) \right) / \sum_i \left(\max_{z \in \mathcal{Y}} z(t_i) - \min_{z \in \mathcal{Y}} z(t_i) \right), \qquad (3)$$

  being $\{t_i\}$ the knots where the curves are observed.

- *Covered proportion*

  $$P_k(y) = \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}\left(\min_{z \in \mathcal{J}_k(y)} z(t_i) \leq y(t_i) \leq \max_{z \in \mathcal{J}_k(y)} z(t_i) \right), \qquad (4)$$

  being T the cardinality of $\{t_i\}$ and $\mathbb{1}(A)$ the indicator function of A.

Table 1 and Figure 5 show the average of these statistics. From the Table 1, we observe 8 outlying curve of 66 possible cases by using Algorithm 1. Meanwhile Algorithm 2 provides bands for all the cases.

Despite this difference, both algorithms are almost on a par. They generate subsamples of similar size where the focal curve is deep, in fact the deepest in half of the cases. From

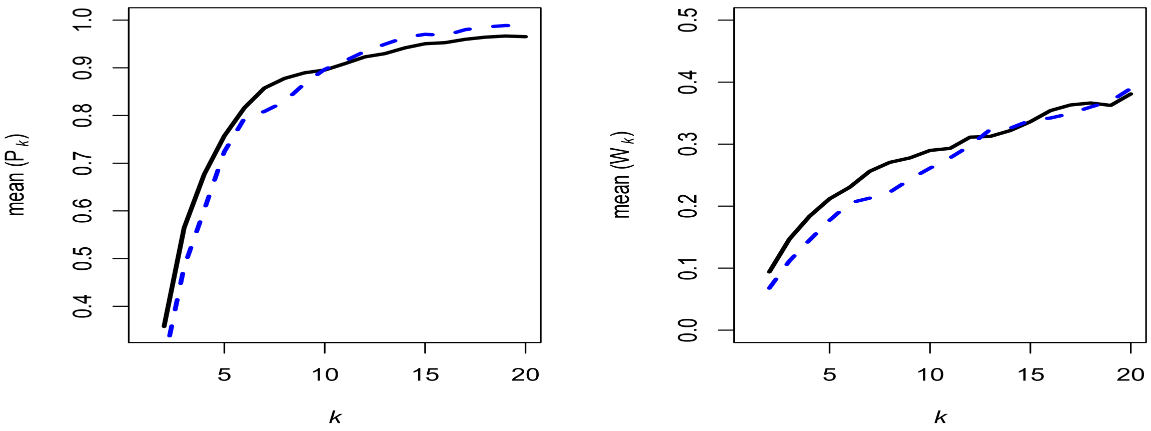

Figure 5: Average of coverage proportion and standardized mean width of the central regions $\mathcal{J}_k$ for monthly sea temperatures. Black line corresponds to Algorithm 1 and dashed blue line to Algorithm 2.

the Figure 5, we note that both algorithms provide central regions $\mathcal{J}_k$ which cover a high proportion of focal curve, from $k \geq 5$. In addition, the regions are delimited by tight bands, up to a 80% thinner than the band that covers the entire sample. This fact produces that the bands preserve the shape of the focal curve.

4 Prediction bands

Now consider that the sample functions $y_1, ..., y_n$ are observed not only on I but on a contiguous interval I_0 where we have not record of y_0. Let $\mathcal{J}(y_0)$ be the subsample for y_0 obtained by restricting the sample functions on I. What we expect is that the shape and magnitud of y_0 is captured by the deepest curves of $\mathcal{J}(y_0)$, not only on I but on I_0. This simple approach can bee seen as a functional version of the nonparametric method introduced by Sugihara and May (1990) for making short-term predictions about the trajectories of quasi-ergodic dynamical processes. It seems intuitively clear that, subject to general conditions on the sample curves, if y_0 is approximately enveloped on I by a set of near curves then they also will surround y_0 on I_0. Next, we explore this idea by simulation.

	Algorithm 1	**Algorithm 2**
outlying curves	5 out of 100	0 out of 100
subsample size	22.031	22.421
depth percentile on $[0,\pi]$	0.987	0.975
depth percentile on $(\pi,2\pi]$	0.987	0.969

Table 2: Average based on 100 random focal curves from the harmonic signals sample separately for the training interval $[0,\pi]$ and the predicting one $(\pi,2\pi]$.

We consider a sample of 100 harmonic signals $y_i(t) = a_i \sin(t) + b_i \cos(t)$, $t \in [0, 2\pi]$, with a_i and b_i being independent and uniformly distributed on $(0, 1)$. These functions are similar to those used by Hyndman and Shang (2010) and Sun and Genton (2011) but with wider range of values of a_i and b_i, allowing more variety of shapes. We simulate 100 focal functions of the same form. Restricting on $I = [0, \pi]$, consider the k-deepest curves $\mathcal{J}_k(y_0)$. We use $I_0 = (\pi, 2\pi]$ for testing our predictions. For illustration, we show in Figure 6 bands with $k = 6$ for two different shapes of the focal curve. The first thing one can notice is that the shape and magnitude of the focal is captured by the bands on both intervals, $[0, \pi]$ and $[\pi, 2\pi]$, no matter which algorithm we are using. The above is confirmed by the results of our simulation. Table 2 shows that the focal remains deep in both intervals and the total number of curves chosen of both algorithms are roughly the same. In Figure 7, we only present the averages for the prediction interval $[\pi, 2\pi]$ since the corresponding curves for averages on $[0, \pi]$ are overlapped. This figure shows that, from $k \geq 5$, the bands delimited by the k deepest curves cover (in mean) the focal one on a large proportion, providing tight prediction bands.

5 Case studies

The results of previous section are not a surprise. Note that the curves considered are determined only by the coefficients a_i and b_i. Therefore, features of neighbour curves to the focal, such as shape and magnitude, are similar on the training interval $(0, \pi)$ and on the predictive interval $(\pi, 2\pi)$. For the same reason, the dimension of these functional data

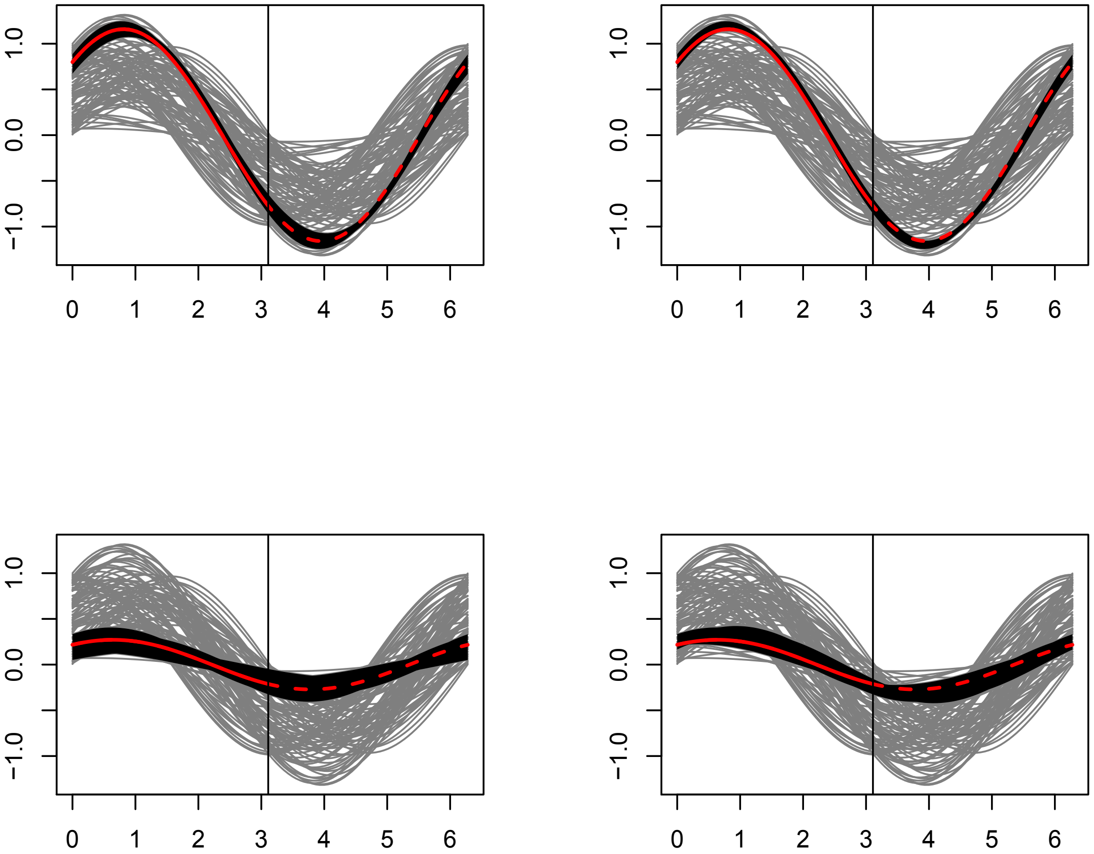

Figure 6: Left-half side of panels: focal curve (solid red line) and band (in black) based on six deepest harmonic curves on $[0, \pi]$ provided by Algorithms 1 (left panels) and 2 (right panels). Right-half side of panels: extended extended bands and focal (dashed red line).

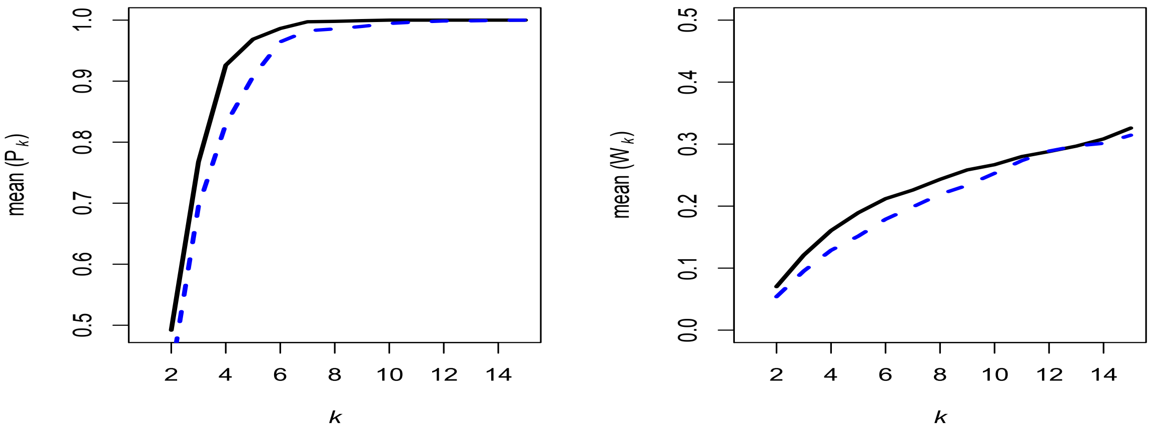

Figure 7: Average over 100 random harmonic signals of coverage proportion and standardized mean width for prediction central regions $\mathcal{J}_k$. Black line corresponds to Algorithm 1 and dashed blue line to Algorithm 2.

can be reduced to two. Hence, both algorithms should be comparable we are not loosing much information by reducing the dimension of the functional data set. Next, we test the algorithms with two real data sets that provides a natural functional setting with more complicated shapes and structures.

5.1 Spanish electricity demand

Data concerning the Spanish electricity demand is available at `http://www.ree.es/es/`, from where we obtained the demand in megawatts (MW) from January first 2014 to December 31st 2016 each 10 minutes. We consider the daily demand as the sample unit, 1096 curves in total. Our prediction exercise consisted in forecasting half day from October first to December 31st of 2016, 92 days, by only using past information. The averages over the 92 predictions of coverage proportions and standardized mean widths of central regions $\mathcal{J}_k$ are shown in Figure 8. We also compute mean depth percentiles and subsample sizes (Table 3).

The bands delimited by the $k \geq 10$ deepest curves are tight (80% smaller in mean than the band produced by the whole sample) and cover the focal function on a high proportion

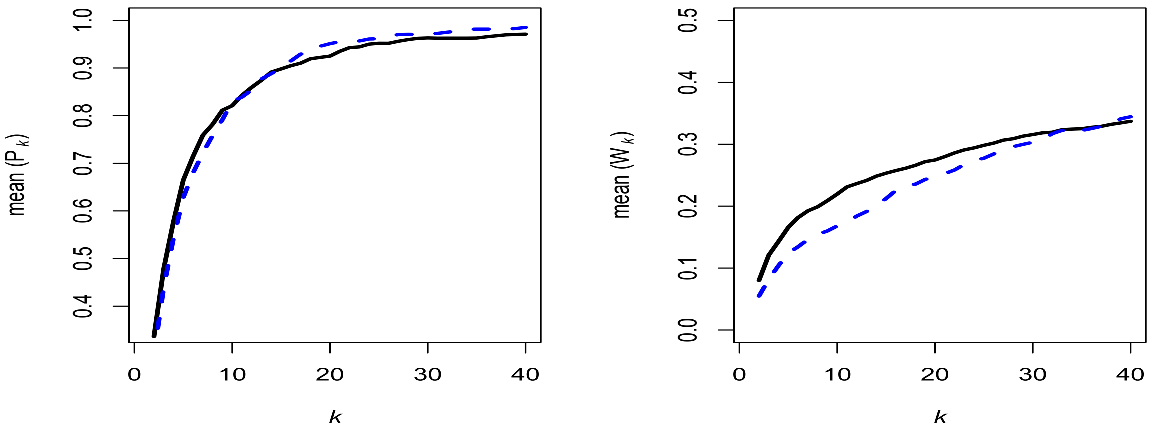

Figure 8: Average of coverage proportion and standardized mean width of $\mathcal{J}_k$ for forecasting daily electricity demand. Black line corresponds to Algorithm 1 and dashed blue line to Algorithm 2.

	Algorithm 1	**Algorithm 2**
outlying curves	1 out of 92	0 out of 92
subsample size	124.362	120.242
depth percentile on $[0, 720]$	0.999	0.985
depth percentile on $(720, 1440]$	0.753	0.725

Table 3: Averages obtained for forecasting electricity demand between October 1 and December 31 of 2016. The training interval consists on the first 720 minutes.

(more than 80% in mean). Although the centrality of the focal function in the interval $(720, 1440]$ decreases with respect to the interval $[0, 720]$, its depth percentile remains high (0.753 and 0.725 for Algorithm 1 and 2 respectively).

As illustration, we show in top panel of Figure 9 the band corresponding to Tuesday, November 29th, 2016 by using Algorithm 2 and $k = 10$. Notably, all the curves used to envelope the focal come from months between November and February, however they correspond to different working days and years.

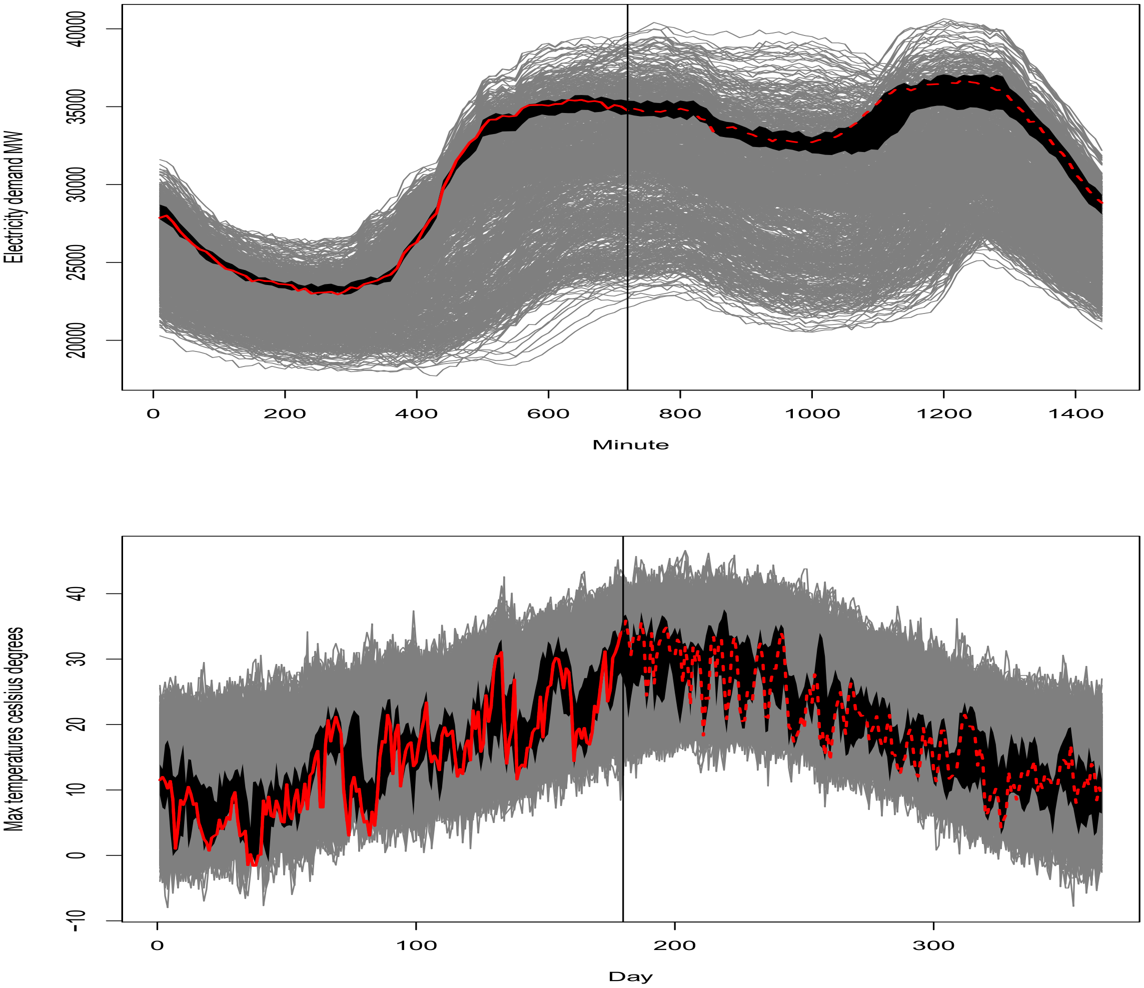

Figure 9: Two examples of predicting bands based on $k = 10$ from Algorithm 2. Top panel: Spanish electricity demand. Focal curve (in red) corresponds to Tuesday, November 29th, 2016. Bottom panel: Spanish daily temperatures. Focal curve corresponds to the station that achieved the minimum temperature of 2015 (Burgos/Villafría).

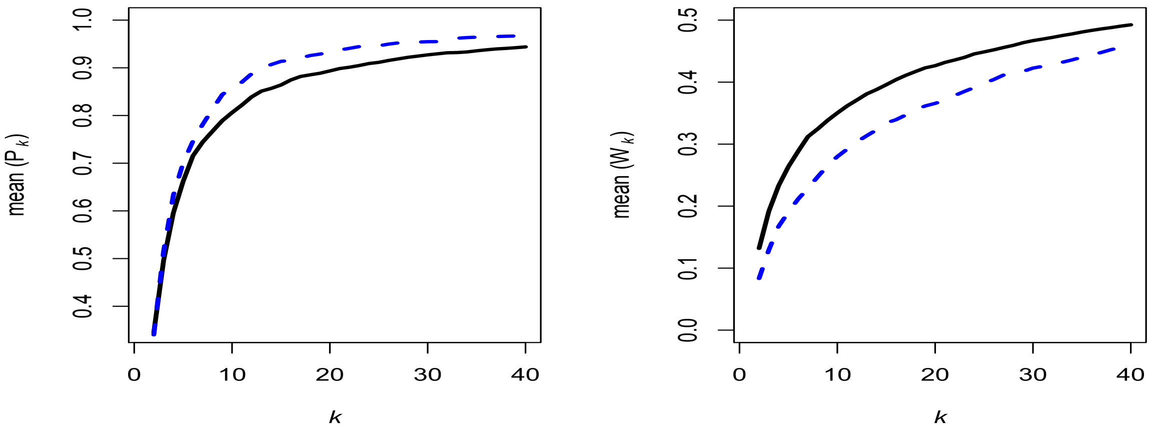

Figure 10: Average of coverage proportion and standardized mean width of $\mathcal{J}_k$ for predicting daily temperatures. Black line corresponds to Algorithm 1 and dashed blue line to Algorithm 2.

5.2 Spanish daily temperatures

Daily maximum temperatures along a year usually show strong oscillations that enrich the shape of the data sets analysed in previous sections. This particularly challenging feature has been resolved in other studies of daily temperatures by averaging (Delicado et al., 2010) or by applying Fourier basis smoothing Ramsay and Silverman (2005). However, in this section we not only considered raw data but the additional difficulty of predicting far horizons. Beyond a mere exercise, the daily register of some weather Spanish stations have large missing intervals. For testing, we consider only complete year. This data set involves years from 1893 to 2015 of the 53 stations located at the capital of provinces and was provided by AEMET (Spanish State Agency of Meteorology). The present study consisted in predict, one by one, the daily temperature from July to December 2015 of the 53 stations, by using the complete sample, excepting the assumed missing interval.

In this example, the bands corresponding to Algorithm1 provides lower mean coverage and then envelope worse the focal in comparison with the results obtained with previous ones. Moreover, the focal decreases considerably in depth from the training interval to the prediction one. We expected these results due to the complexity of the curves that can not be reduced only to the first two principal components. Here, to reduce the dimension

	Algorithm 1	**Algorithm 2**
outlying curves	1 out of 53	0 out of 53
subsample size	216.981	72.711
mean depth percentile on $(0, 182]$	0.995	0.930
mean depth percentile on $(182, 365]$	0.490	0.797

Table 4: Averages obtained for predicting daily temperatures.

implies a lost of information. It is not the case of the bands corresponding to Algorithm 2, which provides comparable results to Algorithm 1 (see Figure 10 and Table 4). We show the band corresponding to $k = 10$ in the bottom panel of Figure 9 for the station that achieved the minimum temperature of 2015 (Burgos/Villafría). Among the curves involved in the band, we observe that there are curves coming from far stations and dated long time ago.

6 Conclusions

In this article, we propose an approach for predicting partially observed functions. The methodology relies on the selection of subsamples that make the function to predict a deep datum in the range of observation. We propose two heuristics for selecting sample curves from a focal function to outwards with the goal of combining proximity and centrality. The bands delimited by the deepest curves of the selected subsample provide tight regions that envelope the unobserved part of the focal, preserving its shape.

The approach presented is nonparametric and phenomenological since it attempts to capture the morphology of the curve to predict without attempting to provide an understanding of the statistical model where the sample come from. Furthermore, the methodology is easy to adapt to any functional sample. We propose two algorithms for implementing our approach. Algorithm 1 is based on the Tukey's depth and the first two robust principal components. Algorthm 2 uses the band depth evaluated on the functional space. The performance of both algorithms is similar for samples where dimensionality reduction does not involve considerable loss of information. In this case, Algorithm 2 seems to be better. This is evident by considering the challenging data set of Spanish daily temperatures. However, both procedures are comparable in terms of performance in our simulation study and in the forecast of the Spanish electricity demand.

The algorithms can easily be extended in several directions. On one hand, for Algorithm 1 we could consider more than two principal components, in order to capture additional features of the sample curves. For this, we only requiere Delaunay tesselations in more than two dimensions. Alternative dimension reduction methods to principal components can be considered, such as other multivariate approaches suggested by Hyndman and Shang (2010) and functional procedures proposed by Shang (2017). On the other hand, although it is true that Euclidean distance between points is the natural way to measure proximity between scores, there are several alternatives to measure nearness between curves. The L^2 norm used in Algorithm 2 can be substituted by other distances, including those that consider shape difference (Marron and Tsybakov, 1995; Minas et al., 2011). Some distances may be more appropriate for detecting certain aspects of the data under study. It is possible also to replace the depth measure used. Nieto-Reyes and Battey (2016) discuss different properties that functional depths should satisfy, being the h-depth proposed by (Cuevas et al., 2007) the one that satisfies more properties. Additionally, one can extend our approach to multivariate functional data, by using surface bands as those discussed by Sun and Genton (2011) and taking advantage of new multivariate functional depth measures (Ieva and Paganoni, 2013; López-Pintado et al., 2014). Finally, other heuristics for selecting sample curves from the focal to outwards could be considered.

References

Aneiros-Pérez, G.rez, G. G., Cao, R., and Vilar-Fernández, J. M. (2011). Functional methods for time series prediction: a nonparametric approach. *J. Forecast.*, 30(4):377–392.

Aneiros-Pérez, G.rez, G. G. and Vieu, P. (2008). Nonparametric time series prediction: A semi-functional partial linear modeling. *Journal of Multivariate Analysis*, 99(5):834–857.

Antoniadis, A., Paparoditis, E., and Sapatinas, T. (2006). A functional wavelet-kernel approach for time series prediction. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(5):837–857.

Arribas-Gil, A. and Romo, J. (2014). Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15(4):603–619.

Chiou, J.-M., Zhang, Y.-C., Chen, W.-H., and Chang, C.-W. (2014). A functional data approach to missing value imputation and outlier detection for traffic flow data. *Transportmetrica B: Transport Dynamics*, 2(2):106–129.

Croux, C. and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1):206–226.

Cuesta-Albertos, J. and Nieto-Reyes, A. (2008). The random tukey depth. *Computational Statistics & Data Analysis*, 52(11):4979–4988.

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

Delaigle, A. and Hall, P. (2013). Classification using censored functional data. *Journal of the American Statistical Association*, 108(504):1269–1283.

Delicado, P., Giraldo, R., Comas, C., and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics*, 21(3-4):224–239.

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

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. 10(2):419–440–.

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

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

James, G., Hastie, T., and Sugar, C. (2000). Principal component models for sparse functional data. *Biometrika*, 87(3):587–602.

James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):533–550.

Jones, M. C. and Rice, J. A. (1992). Displaying the important features of large collections of similar curves. *The American Statistician*, 46(2):140–145.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, 27(3):783–840.

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

López-Pintado, S. and Romo, J. (2011). A half-region depth for functional data. *Computational Statistics & Data Analysis*, 55(4):1679–1695.

López-Pintado, S., Romo, J., and Torrente, A. (2010). Robust depth-based tools for the analysis of gene expression data. *Biostatistics*, 11(2):254–264.

López-Pintado, S., Sun, Y., Lin, J. K., and Genton, M. G. (2014). Simplicial band depth for multivariate functional data.

Marron, J. S. and Tsybakov, A. B. (1995). Visual error criteria for qualitative smoothing. *Journal of the American Statistical Association*, 90(430):499–507.

Minas, C., Waddell, S. J., and Montana, G. (2011). Distance-based differential analysis of gene curves. *Bioinformatics*, 27(22):3135–3141.

Narisetty, N. N. and Nair, V. N. (2015). Extremal depth for functional data and applications. *Journal of the American Statistical Association*, pages 1–38.

Nieto-Reyes, A. and Battey, H. (2016). A topologically valid definition of depth for functional data. pages 61–79.

Okabe, A., Boots, B., Sugihara, K., Chiu, S. N., and Kendall, D. G. (2008). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams.* John Wiley & Sons, Inc.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis.* Springer, New York, 2nd edition.

Shang, H. L. (2017). Functional time series forecasting with dynamic updating: An application to intraday particulate matter concentration. *Econometrics and Statistics*, 1:184–200.

Shang, H. L. and Hyndman, R. (2011). Nonparametric time series forecasting with dynamic updating. *Mathematics and Computers in Simulation*, 81(7):1310–1324.

Sugihara, G. and May, R. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 344(6268):734–741.

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

Terrell, G. R. and Scott, D. W. (1992). Variable kernel density estimation. *The Annals of Statistics*, 20(3):1236–1265.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematics (Vancouver, 1974)*, volume 2, pages 523–531.

Yao, F., Muller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.