Localization processes for functional data analysis

Antonio Elías[1], Raúl Jiménez[2] and J. E. Yukich[2,3*]

[1]OASYS Group, Department of Applied Mathematics, Universidad de Málaga, Spain.
[2]Department of Statistics, Universidad Carlos III de Madrid, Spain.
[3*]Department of Mathematics, Lehigh University, USA.

*Corresponding author(s). E-mail(s): joseph.yukich@lehigh.edu;
Contributing authors: aelias@uma.es; rjjimene@est-econ.uc3m.es;

Abstract

We propose an alternative to $\boldsymbol{k}$-nearest neighbors for functional data whereby the approximating neighboring curves are piecewise functions built from a functional sample. Using a locally defined distance function that satisfies stabilization criteria, we establish pointwise and global approximation results in function spaces when the number of data curves is large. We exploit this feature to develop the asymptotic theory when a finite number of curves is observed at time-points given by an i.i.d. sample whose cardinality increases up to infinity. We use these results to investigate the problem of estimating unobserved segments of a partially observed functional data sample as well as to study the problem of functional classification and outlier detection. For such problems our methods are competitive with and sometimes superior to benchmark predictions in the field. The R package `localFDA` provides routines for computing the localization processes and the estimators proposed in this article.

Keywords: Nearest neighbors, Incomplete observations, Outlier detection

1 Introduction

The k-nearest neighbors (kNN) method has been identified by IEEE as one of the top algorithms for solving multivariate statistical problems on large datasets (Wu et al., 2008). It is particularly useful for classification and regression, where the method is based on the idea that similar patterns must belong to the same class and near explanatory variables will have similar response variables. Among other applications of kNN in the multivariate setting, we also include clustering (Brito et al., 1997), outlier detection (Ramaswamy et al., 2000) and time series forecasting (Martínez et al., 2017). Beyond its effectiveness, the popularity of the method is due in part to its conceptual ease and implementation. This has sparked the interest of researchers, who over many years have developed not only applications but also the mathematical theory, making the method an essential tool in nonparametric multivariate statistics. A critical review of the seminal literature on the asymptotic theory related to the application of the method to classification, regression and density estimation is provided by Chapter 6 of Györfi et al. (2002).

In the context of functional data analysis (FDA), the kNN method has also been explored. For example, both Zhang et al. (2010) and Elías et al. (2022) address the problem of forecasting functional time series via functional kNN and Hubert et al. (2017) constructs classifiers for functional data also based on kNN. However, the asymptotic theory of these methods has remained undeveloped until now. Asymptotic results of methods based on functional kNN are mainly related to regression estimation when the response variable has finite dimension (Biau et al., 2010; Kudraszow and Vieu, 2013). Some of these results have been extended to other operations on the response variable (Kara et al., 2017), such as conditional distribution and conditional hazard function. And, exceptionally, Lian (2011) proves the consistency of some regression estimates based on kNN when both dependent and independent variables are functions. Thus, in the functional data setting, the mathematical theory of the kNN rule is relatively unexplored.

Our first purpose is to fill this lacuna and to provide the asymptotic theory for methods either based on or inspired by kNN. The proofs of the main results rely on the theory of stabilizing functionals. This theory has been mainly developed for establishing limit theorems for statistics arising in stochastic geometry (Schreiber, 2010). At its core, this theory is applicable to statistics which are expressible as sums of score functions which depend on local data in a well-defined way. Statistics involving multivariate kNN are prime examples of locally defined score functions. In this context we introduce a kth localization process which well approximates a target process in both a pointwise and L^1 sense. We exploit the fact that the kth localization process is locally defined to rigorously develop its first and second order limit theory.

Our second purpose is to review some problems of the FDA literature from a perspective enriched by the new asymptotic results. In particular, we consider the problem of estimating unobserved values of a partially observed functional data sample, a question prominently addressed in the literature (Kraus, 2015; Yao et al., 2005). As has been already reported (Zhang et al., 2010), the kNN method is a natural approach for addressing this problem. We carried out a comparative study in the reconstruction problem on data consisting of yearly curves of daily Spanish temperatures and also on data consisting of Japanese age-specific mortality rates. The proposed kNN method is shown to provide an estimator having relatively small MSE when compared with competing methods. The estimator often provides estimates which are superior to benchmark predictions in the field and is shown to be consistent under some regularity conditions. We also consider the problems of classification and outlier detection. Specifically, we introduce a probabilistic functional classifier inspired by the kNN rule. The classification method is based on the asymptotic normality of an empirical distance which only considers a finite number of sample curves but takes advantage of the fact that the data live in a function space, giving rise to empirical methods in such spaces. The classification method proposed here is compared with standard ones on three datasets, including fighter plane datasets, the Berkeley growth study dataset, as well as a fat absorbance dataset. As far as we know, this type of asymptotic result is new and is particular to the functional setting. The problem of outlier detection is straightforwardly tackled by considering one-class classification. We carry out a comparative study with other widely used methods in FDA (Arribas-Gil and Romo, 2014; Li et al., 2012; Sun and Genton, 2011) which shows that our approach performs well across different test datasets, including the dataset of Japanese age-specific mortality rates, in which it is shown that the localization processes detect atypical features of the data where other methods do not. In addition, with the new classifier we may predict classification probabilities rather than only outputting the most likely class.

1.1 Definitions and terminology

Functional data are typically viewed as independent realizations of a stochastic process with smooth trajectories observed on a compact interval (Yao et al., 2005). Consequently, we consider a stochastic process $X = \{X(t) : t \in [a, b]\}$ with continuous sample paths. Following standard practice we set $[a, b] = [0, 1]$. Let $X_1, \ldots, X_n$ be independent copies of the process $X = X(t), t \in [0, 1]$. The $X_i, i \geq 1$, take values in $C[0, 1]$, the space of continuous functions on $[0, 1]$. The classical way of defining a distance between sample paths X_i and X_j is to use the p-norm, $p \geq 1$, also called the Minkowski distance,

$$D(X_i, X_j) = \Big(\int_0^1 |X_i(t) - X_j(t)|^p \Big)^{1/p} dt.$$

The Hausdorff and other distances may be defined in terms of distances between two nonaligned points of the curves X_i and X_j. By two nonaligned points, we mean $(t, X_i(t))$ and $(s, X_j(s))$ with $t \neq s$. Such distances are not considered in the present work. In any case, given a distance $D(\cdot, \cdot)$ between two functions, the global nearest neighbor to X_i is defined by

$$X_i^{(1)} = \arg \min_{\{X_j : j \neq i\}} D(X_i, X_j).$$

Iterating, for $k \in \{2, ..., n-1\}$, the global kNN to X_i is defined as the nearest neighbor in the subsample $\{X_1, \dots, X_n\} \setminus \{X_i^{(1)}, \dots, X_i^{(k-1)}\}$.

In this paper we consider *local* nearest neighbors to a given curve which without loss of generality is taken to be X_1. We begin by considering the nearest function to X_1 in the pointwise sense. This gives rise to the stochastic process

$$\hat{X}_1^{(1)}(t) = \hat{X}_{n,1}^{(1)}(t) = \arg \min_{\{X_j(t) : j \neq 1\}} |X_1(t) - X_j(t)|, \quad t \in [0, 1],$$

which we call the *first localization process.* This process consists of a union of sample curve segments which are the nearest sample observations to X_1. We define the k-nearest sample piecewise function to X_1 by iterating, in way similar to how we defined the global kNN: First, for $t \in [0, 1]$, let $G_1^{(1)}(t) = \{X_j(t) : j \neq 1\}$. Then, for $2 \leq k \leq n-1$, recursively define the *kth localization process* by

$$\hat{X}_1^{(k)}(t) = \hat{X}_{n,1}^{(k)}(t) = \arg \min_{x(t) \in G_1^{(k)}(t)} |X_1(t) - x(t)|, \quad t \in [0, 1], \tag{1}$$

where

$$G_1^{(k)}(t) = G_1^{(k-1)}(t) \setminus \{\hat{X}_1^{(k-1)}(t)\}.$$

This curve is the central object of our studies. We will show that it well approximates X_1 in a pointwise and global sense. Though X_1 is continuous by assumption, the process $\hat{X}_1^{(1)}$ need not be continuous and, in general, may have a finite number of discontinuities. The same remark applies to $\hat{X}_1^{(k)}$.

The localization distances between $X_1(t)$ and $\hat{X}_i^{(k)}(t), t \in [0, 1]$, give rise to the *$k$th localization width process*

$$L_n^{(k)}(t) = L^{(k)}(X_1(t), \{X_j(t)\}_{j=1}^n) = |X_1(t) - \hat{X}_1^{(k)}(t)|, \ t \in [0, 1].$$

The R package `localFDA` (Elías et al., 2021) provides programs for computing the localization and localization width processes. Under mild conditions on the marginal densities of $X(t), 0 \leq t \leq 1$, we shall show that the L^1 norm of $L_n^{(k)}(\cdot)$ is $O\big(k/2n\big)$. Hereinafter, we assume that the marginal probability density of $X(t)$, denoted by κ_t, exists for almost all $t \in [0, 1]$. Denote by $S(\kappa_t)$ the possibly unbounded support of κ_t.

We will assess the closeness of the kth localization process $\hat{X}_1^{(k)}(t), t \in [0,1]$, to the data curve $X_1(t), t \in [0,1]$, by studying the *re-scaled localization width process*

$$W_n^{(k)}(t) = W^{(k)}\left(X_1(t), \{X_j(t)\}_{j=1}^n\right) = \frac{2n}{k} L_n^{(k)}(t), \quad t \in [0,1]. \tag{2}$$

The re-scaled localization distance $W_n^{(k)}$ is invariant under any affine transformation of the data, i.e., transformations of the data by functions of the type $T(x) = ax + b$, with a, b scalars, leave $W_n^{(k)}$ unchanged. Both the results and methods discussed in this paper are invariant under affine transformations.

1.2 Outline of this work

This paper is organized as follows. Section 2 presents three types of asymptotic results.

(a) *Pointwise convergence.* We show mean and distributional convergence of $W_n^{(k)}(t)$, at a fixed $t \in [0,1]$, when $n \to \infty$.
(b) *Process convergence.* Under regularity conditions on the density of the data, the L^1 norm of the average difference between $W_n^{(k)}(t), t \in [0,1]$, and a limit localization process is $o(1)$. This yields that the expected L^1 norm of $|X_1(t) - \hat{X}_1^{(k)}(t)|$ is $O(k/2n)$.
(c) *Asymptotic normality.* For a fixed number of curves n, fixed k, and for $\mathcal{T}_m$ a set of m i.i.d. locations, the so-called empirical localization distance defined by $m^{-1} \sum_{t \in \mathcal{T}_m} L_n^{(k)}(t)$, follows a Gaussian distribution when m increases up to infinity.

Limit results (a) and (b) are used in Section 3, where we consider estimation of missing values of partially observed data via kNN. Limit result (c) is used in Section 4, which provides a probabilistic classifier and a outlier detection tool based on localization distances. A general discussion of both theoretical and practical results is given in Section 5 whereas the Appendix A provides the proofs of our main results of Section 2.

2 Asymptotics results for localization processes

This section is devoted to providing asymptotic theory for localization processes. Asymptotic results for global nearest neighbors as well as a description of the statistical methods introduced in this paper are discussed in later sections.

2.1 Asymptotics for localization processes with large data sizes

We first assess the pointwise behavior of the re-scaled localization width process on large data sets. Let $|A|$ denote the Lebesgue measure of the set A.

Theorem 2.1. *For $k \in \mathbb{N}$ and almost all $t \in [0,1]$, we have*

$$\lim_{n\to\infty} \mathbb{E}W_n^{(k)}(t) = |S(\kappa_t)| \in (0,\infty] \tag{3}$$

and

$$\lim_{n\to\infty} \mathrm{Var}[W_n^{(k)}(t)] = \left(1+\frac{1}{k}\right)\int_{S(\kappa_t)} \frac{1}{\kappa_t(y)}dy - |S(\kappa_t)|^2 \in \left[\frac{|S(\kappa_t)|^2}{k},\ \infty\right]. \tag{4}$$

In addition, as $n \to \infty$,

$$W_n^{(k)}(t) \xrightarrow{\mathcal{D}} W_\infty^{(k)}(t) = \frac{2}{k}\frac{1}{\kappa_t(X_1(t))}\Gamma(k,2), \tag{5}$$

where $\Gamma(k,2)$ is a Gamma random variable with shape parameter k and scale parameter 2.

The Jensen inequality

$$\phi\left(\frac{1}{b-a}\int_a^b f(x)dx\right) \le \frac{1}{b-a}\int_a^b \phi(f(x))dx$$

with $\phi(x) = x^{-1}, x > 0$, shows that $\frac{1}{|S(\kappa_t)|}\int_{S(\kappa_t)} \frac{1}{\kappa_t(y)}dy \ge |S(\kappa_t)|$. Thus the asympotic variance in (4) is greater than or equal to $\frac{1}{k}|S(\kappa_t)|^2$ and attains equality when κ_t is the uniform density on $S(\kappa_t)$.

We next assess the global behavior of the localization process in the L^1 norm on $C[0,1]$. We find conditions under which the L^1 norm of the expected difference between the localization width process $W_n^{(k)}(t)$ and $W_\infty^{(k)}(t)$ defined at (5) converges to zero. Thus, the target function X_1 is *globally* well approximated by the localization process. Specifically, the asymptotic expected L^1 error in locating a typical curve by its kth localization process is $O(\frac{k}{2n}\int_0^1 |S(\kappa_t)|dt)$, which depends on k and on the Lebesgue measure of the support of the underlying distribution of the data. This happens provided that the data is *regular from below*, i.e., for almost all $t \in [0,1]$ we have $|S(\kappa_t)| < \infty$ and there exists $\kappa_{\min} > 0$ and an interval $S_\delta \subset S(\kappa_t)$, with $|S_\delta| = \delta > 0$, such that

$$\inf_{x\in S_\delta} \kappa_t(x) \ge \kappa_{\min}. \tag{6}$$

Examples of data which are regular from below include harmonic signals of the form $X(t) = A\sin(2\pi t) + B\cos(2\pi t)$, where the coefficients A and B are independent random variables having a uniform distribution on some compact interval. These processes have been used in simulation studies by several authors; e.g. by Hyndman and Shang (2010) and Sun and Genton (2011), among others. More generally, finite Fourier sums with independent random coefficients having a probability density defined on a compact interval are also examples of processes satisfying (6). Due to either physical or biological restrictions, or to limitations of supply, functional data are often bounded. This includes, for example, mortality, fertility and migration rates (Hyndman and Ullah, 2007; Hyndman and Booth, 2008); curves of surface air temperatures and precipitation (Dai and Genton, 2018); electricity market data (Liebl, 2019) and functional data from medical studies (Kraus, 2015; Yao et al., 2005). For such data, we may assume the data is regular from below. The next two process level results apply to such data curves.

Theorem 2.2. *Assume that the data is bounded and regular from below as at* (6)*. Then for* $k \in \mathbb{N}$

$$\lim_{n\to\infty} \int_0^1 \mathbb{E}|W_n^{(k)}(t) - W_\infty^{(k)}(t)|dt = 0. \tag{7}$$

Consequently $\lim_{n\to\infty} \int_0^1 \mathbb{E}W_n^{(k)}(t)dt = \int_0^1 |S(\kappa_t)|dt$ *and the average* L^1 *error satisfies*

$$\mathbb{E}\int_0^1 |X_1(t) - \hat{X}_1^{(k)}(t)|dt \leq \frac{k}{2n}\left(o(1) + \int_0^1 |S(\kappa_t)|dt\right) = O\left(\frac{k}{2n}\right). \tag{8}$$

Under further conditions we may extend the convergence (8) to curves other than $\hat{X}_1^{(k)}$. This will be spelled out in more detail in Proposition 1 in Section 3. To prepare for this, we consider the case when $k = k(n)$ increases with n. This result is used in Section 3 to aid in reconstructing curves with missing data.

Theorem 2.3. *Assume that the data is bounded and regular from below as at* (6)*. Assume* $\kappa_t, t \in [0, 1]$*, is* α*-Hölder continuous, i.e., there is* $\alpha \in (0, 1]$ *such that*

$$|\kappa_t(x) - \kappa_t(y)| \leq C|x - y|^\alpha, \quad x, y \in S(\kappa_t).$$

Let $k = k(n)$ *satisfy* $\lim_{n\to\infty} \frac{k^{1+\alpha}}{n^\alpha} = 0$*. Then for almost all* $t \in [0, 1]$

$$\lim_{n\to\infty} \mathbb{E}W_n^{(k)}(t) = |S(\kappa_t)| \in (0, \infty] \tag{9}$$

and

$$\lim_{n\to\infty} \mathrm{Var}[W_n^{(k)}(t)] = \int_{S(\kappa_t)} \frac{1}{\kappa_t(y)}dy - |S(\kappa_t)|^2 \in [0, \infty]. \tag{10}$$

Additionally, if κ_t is α-Hölder continuous for almost all $t \in [0,1]$, then

$$\lim_{n\to\infty} \int_0^1 \mathbb{E}W_n^{(k)}(t)dt = \int_0^1 |S(\kappa_t)|dt. \tag{11}$$

The right-hand side of (10) vanishes if κ_t is uniform on its support. In this case $W_n^{(k)}(t)$ converges to $|S(\kappa_t)|$ in probability as $n \to \infty$.

2.2 Stochastic behavior of empirical localization distances

One often observes functional data on a discrete time set $\{t_1, \ldots, t_m\}$. In such cases, the average distance between X_1 and its kth localization process (1) is a global measure of nearness. This distance is given by the average width

$$\frac{1}{m}\sum_{r=1}^{m} L^{(k)}(X_1(t_r), \{X_j(t_r)\}_{j=1}^n). \tag{12}$$

Here we focus on the distributional behavior of (12) when $\{t_1, \ldots, t_m\}$ is the realization of i.i.d. uniform random variables $T_1, ..., T_m$ on $[0,1]$. This gives rise to an empirical localization width process and goes as follows. We fix n, the number of data curves. We evaluate the localization distance with respect to X_1 at each $T_r, 1 \le r \le m$. This generates the so-called *empirical localization distance* between X_1 and its kth localization process namely

$$\frac{1}{m}\sum_{r=1}^{m} L^{(k)}(X_1(T_r), \{X_j(T_r)\}_{j=1}^n) = \frac{1}{m}\sum_{r=1}^{m} L_n^{(k)}(T_r). \tag{13}$$

The empirical localization distance is simply the localization width $L_n^{(k)}(\cdot)$ averaged over the sample $\{T_r\}_{r=1}^m$.

The localization widths $L_n^{(k)}(T_r)$, $1 \le r \le m$, exhibit dependence in general. However, if $L_n^{(k)}(t')$ depends only on the values of $L_n^{(k)}(t)$ at preceding data points $t \in \{T_r\}_{r=1}^m$ within distance $\frac{M}{\sqrt{m}}$ of t', where M is a fixed positive constant, then the asymptotic normality as $m \to \infty$ of the empirical localization distance follows from M-dependence as follows. Such a dependency assumption holds if the data has a Markovian structure, with $X(t)$ depending only on the immediate past $X(t^-)$.

Theorem 2.4. *Fix n, the number of data functions. Let $k \in \{1, ..., n-1\}$. Assume that the localization distances $L_n^{(k)}(t), t \in \{T_r\}_{r=1}^m$, depend only on the* *values of $L_n^{(k)}(\cdot)$ at preceding data points $t \in \{T_r\}_{r=1}^m$ within distance $\frac{M}{\sqrt{m}}$ of t, where M is a fixed positive constant. Then as $m \to \infty$ we have*

$$\frac{\sum_{r=1}^m (L_n^{(k)}(T_r) - \mathbb{E}L_n^{(k)}(T_r))}{\sqrt{\mathrm{Var}[\sum_{r=1}^m L_n^{(k)}(T_r)]}} \xrightarrow{\mathcal{D}} N(0,1). \tag{14}$$

The expected value and the variance in (14) may be estimated by sample means and sample variances of empirical localization distances. To achieve this, we must observe the empirical localization distance to each sample curve, not only to X_1. We compute the empirical localization distance to X_i by replacing X_1 with X_i in (13). Let $L_i^{(k)}$ be the corresponding statistics. The sample mean and sample variance are

$$\bar{L}^{(k)} = \frac{1}{n}\sum_{i=1}^n L_i^{(k)} \quad \text{and} \quad S_L^2 = \frac{1}{n-1}\sum_{i=1}^n (L_i^{(k)} - \bar{L}^k)^2.$$

When n is large, $T_i^{(k)} = (L_i^{(k)} - \bar{L}^{(k)})/S_L$ is approximately distributed as the centered and normalized empirical localization distance on the left-hand side of (14) . Thus, Theorem 2.4 suggests that the statistics $T_i^{(k)}$ could be used for testing whether X_i *is properly localized by the data* in accordance with an underlying Gaussian distribution. We explore this idea for classification and outlier detection in Section 4.

3 Reconstruction of partially observed data via *k*NN

In this section, we first use localization processes for establishing consistency conditions of estimators based on global kNN for the reconstruction of partially observed functional data. Second, we implement the functional kNN estimator in Subsection 3.1 and apply the method in two case studies in Subsection 3.2. The performance is compared with the estimates obtained from three benchmark methods.

We are interested in the global proximity of estimators based on kNN, achieved in part by Proposition 1 below, which relies upon the asymptotics results of Section 2, particularly Theorem 2.3. In the first part of this section, for each fixed $j \in \{1, ..., n-1\}$, we find an estimator $X^{(n,j)}$ which is the jth nearest neighbor to X_1 with respect to the L^2 distance on the set where X_1 is observed. We use Theorem 2.3 to show that $X^{(n,j)}$ well approximates X_1 everywhere. A convex combination of $X^{(n,j)}, j \in \{1, ..., n-1\}$, provides a functional kNN estimator which is used in Subsection 3.2 to reconstruct missing data from yearly curves of Spanish temperatures, as well as for curves of mortality data.

3.1 The functional kNN estimator

One might hope that two curves which are near on a set $S \subset [0,1]$ and which are copies of the same process should remain near on $[0,1] \setminus S$. In particular, if S is blindly chosen and with large Lebesgue measure, one could hope to achieve this proximity without taking into account prior morphological information, such as shape and complexity, of the sample curves. Here we show that this turns out to be the case, subject to mild assumptions on the data. This is achieved by making use of k-nearest neighbor methods for reconstructing partially observed data.

As is customary in the literature, we model partially observed functional data by considering a random mechanism Q that generates compact subsets of $[0,1]$. These sets correspond to ranges where sample paths are observed. Formally, $O_1, \dots, O_n$ are independent random closed sets from Q such that X_i is observed on O_i and is missed on $M_i = [0,1] \setminus O_i$. We also will assume data are Missing-Completely-At-Random, i.e., the sets $\{O_i\}$ are independent of the sample paths (Kneip and Liebl, 2020). Without loss of generality, suppose also that there is no time which is almost surely censured. That is to say we assume $\mathbb{P}(O_i \text{ contains } s) > 0$ for almost all $s \in [0,1]$.

To simplify the notation, consider first the case in which just one sample path is partially observed, say X_1. Therefore we assume for now that $X_2, \dots, X_n$ are fully observed on $[0,1]$. Instead of the Minkowski distances to X_1 taken on the complete observation range $[0,1]$, we now consider such distances restricted to O_1, namely for $p \in \{1,2\}$ we put

$$D_p(X_j) = \left(\int_{O_1} |X_j(t) - X_1(t)|^p \right)^{1/p} dt, \; 2 \leq j \leq n-1. \tag{15}$$

For $1 \leq j \leq n-1$, denote by $X^{(n,j)}$ the jNN to X_1 with respect to this distance, where we suppress the dependence of $X^{(n,j)}$ on p. For simplicity of exposition we shall fix $p = 2$ in the definition of $X^{(n,j)}$. To estimate X_1 on M_1, we adopt the kNN methodology and consider estimators based on convex combinations of $X^{(n,j)}$ with the form

$$\hat{X}^{(n)}_{k\mathrm{NN}} = \sum_{j=1}^{r} w_j \cdot X^{(n,j)}, \tag{16}$$

with $w_j > 0$, for $1 \leq j \leq k$, and $\sum_{j=1}^{r} w_j = 1$. The choice of r and suitable weights $\{w_j\}_{j=1}^{r}$ will be discussed later. We start by providing conditions for the consistency of this estimator.

Choose an arbitrary $l \neq 1$ and consider the random interval on which $X_l(t)$ is closer to $X_1(t)$ than is the kth localization process $\hat{X}_1^{(k)}(t)$ defined at (1).

That is to say we consider the random interval

$$I^{(k)}(X_l) = \Big\{ t \in [0,1] : |X_l(t) - X_1(t)| \le |\hat{X}_1^{(k)}(t) - X_1(t)| \Big\}. \tag{17}$$

Note $\mathbb{P}(I^{(k)}(X_l)$ contains $s) = k/(n-1)$ for all $s \in [0,1]$. Since $X^{(n,j)}$ is selected by its proximity to X_1 on O_1, and since O_1 is independent of X_1 and $\hat{X}_1^{(k)}$, one expects, as k increases up to $n-1$, that $\mathbb{P}(I^{(k)}\big(X^{(n,j)}\big)$ contains $s)$ increases up to 1 faster than $\mathbb{P}(I^{(k)}(X_l)$ contains $s)$ for any fixed j and $s \in [0,1]$. More formally, we will consider the following assumption:

$$\lim_{n\to\infty} \mathbb{P}\big(I^{(k)}\big(X^{(n,j)}\big) \text{ contains } s\big) = 1 \quad \text{for some} \quad k = o(\sqrt{n}), \quad s \in [0,1]. \tag{18}$$

Given the features of many functional data used in practice, condition (18) does not appear too unusual. In many cases, the functions are smoothed data arising from Fourier analysis. This is the reason why simulation studies often consider Fourier sums with random coefficients for generating test data. In this context we note that Fourier sums are close to a target curve whenever the respective coefficients are close. Moreover, if the Fourier sums are close to a target on an observable window in $[0,1]$, then they are close everywhere in $[0,1]$, since the coefficients do not depend on t, which in general is not the case for data arising from wavelet or spline analysis. In the case of Fourier sums, if j is fixed the jNN is on average at distance $O(1/n)$ for each $t \in [0,1]$. On the other hand, if $k = o(\sqrt{n})$, the kth localization process is at a distance $O(k/n) = o(1/\sqrt{n})$, verifying (18). For verifying the above, we estimated $\mathbb{P}\big(I^{(k)}\big(X^{(n,j)}\big)$ contains $s\big)$ based on 1000 replicates of (X_1, O_1) under these three conditions: (1) $n = 2500$ and k ranges over the integers up to 250, (2) O_1 is the set obtained by removing at random one of the three closed intervals of the subdivision of $[0,1]$ induced by two independent Uniform(0,1) random variables and (3) X_i is a linear combination of sines and cosines with independent normal coefficients, as those used for generating data in previous studies (Kneip and Liebl, 2020; Kraus, 2015). The results are summarized in Figure A1 of Appendix A.

The estimator $\hat{X}_{k\mathrm{NN}}^{(n)}$ is consistent under regularity conditions, as seen by the next result.

Proposition 1. *Assume the data is bounded and regular from below as at* (6). *Assume for almost all $t \in [0,1]$ that κ_t is Lipschitz and that (18) holds. Then*

$$\lim_{n\to\infty} \mathbb{P}\Big(\int_0^1 |\hat{X}_{kNN}^{(n)} - X_1(t)| dt < \varepsilon \Big) = 1.$$

Proof. See Appendix A.

We follow the previous literature (Hubert et al., 2017; Zhang et al., 2010) and consider Minkowski distances with $p = 1$ and $p = 2$. In the forecasting context, both functional data and univariate time series, the weights for these Minkowski distances have been already suggested (Zhang et al., 2010; Martínez et al., 2017). We use these recommendations and set $w_j = D_p\big(X^{(j)}\big)^{-p} / \sum_{i=1}^{r} D_p\big(X^{(i)}\big)^{-p}$, $D_p(\cdot)$ being the distance to X_1 defined in (15). The value of r used to define the kNN estimator is chosen by minimizing the mean squared error between the estimator and the target X_1 on the observation range. This is

$$r = \arg\min_{r} \int_{O_1} |\hat{X}^{(n)}_{k\text{NN}}(t) - X_1(t)|^2 dt.$$

The Mean Squared Errors on M_1 (MSE), that is to say $\int_{M_1} |\hat{X}(t) - X_1(t)|^2 dt$, are used for evaluating the performance of the estimator.

As is customary, we suppose there is a proportion of curves completely observed. In this case, we may repeat the above approach for estimating any sample curve partially observed by choosing its jNN from among the curves which are fully observed.

To illustrate the method, we conducted a simulation study based on two real case studies. The differences between the results obtained by the functional kNN method based on the Minkowski distance with $p = 1$ and $p = 2$ were negligible, being slightly superior for $p = 2$. For the purposes of summarizing the data, we only report results for $p = 2$.

3.2 Reconstruction study

We apply the functional kNN estimator (16) to real data and compare its performance with benchmark methods in the literature. Case studies involving yearly curves of daily Spanish temperatures and Japanese age-specific mortality rates are considered.

3.2.1 Yearly curves of daily Spanish temperatures

Yearly curves of daily temperatures are common in FDA (Dai and Genton, 2018; López-Pintado and Romo, 2009; Ramsay and Silverman, 2005). We consider 2786 such curves from 73 weather stations located in the capital cities of 50 Spanish regions (provinces). The data was obtained from http://www.aemet.es/, the Meteorological State Agency of Spain (AEMET) website. The date at which the data was first recorded varies from station to station. For example, the Madrid-Retiro station reports records from 1893 onwards whereas the Barcelona-Airport started in 1925 and Ceuta from 2003. On the other hand, it is likely that some states failed at some moment to record data. The point is that there are several incomplete years (Febrero-Bande et al., 2019). With the aim of estimating the missing data, we test

	kNN	KRAUS	KL20	PACE
Spanish Temperature	**0.1351** **(0.1134)**	7.4478 (1.7490)	3.4935 (10.487)	4.0523 (0.3510)
Japanese Mortality	**0.0197** **(0.0164)**	0.0980 (0.0554)	2.5640 (1.1701)	6.6073 (3.4661)

Table 1 Mean running time in seconds observed from 1000 reconstruction exercises based on yearly curves of Spanish daily temperatures and Japanes age-specific mortality rates. Standard deviations are between parentheses.

several methods with a simulation study based on this data set. From the 2786 fully observed curves we selected one at random, labeled as X_1, at which we censored a random number of consecutive days. Term O_1 represents the uncensored days. The average number of censured days was 122, a third of the year. We repeat this procedure 1000 times and estimate the censured data by using the following methods, which we append acronyms so as to easily refer to them in what follows:

1. The kNN estimator described above.
2. KRAUS (Kraus, 2015). This is a functional linear ridge regression model for completing functional data based on principal component analysis. KRAUS estimates scores of partially observed functions by estimating their best predictions as linear functionals of the observed part of the trajectory. Then KRAUS uses a functional completion procedure that recovers the missing piece by using the observed part of the curve. The code was obtained from the author's website (https://is.muni.cz/www/david.kraus/web_files/papers/partial_fda_code.zip).
3. KL20 (Kneip and Liebl, 2020). This reconstruction method belongs to a new class of functional operators which includes the classical regression operators as a special case. The code was obtained from the author's repository (https://github.com/lidom/ReconstPoFD).
4. PACE (Yao et al., 2005). This is the most cited nonparametric method to impute missing data of sparse longitudinal data. The method is based on estimations of the classical eigenfunctions, eigenvalues and scores of truncated Karhunen-Loève decompositions. For implementation, we used the code from the package `fdapace` (Chen et al., 2020).

By far the best method was kNN. For ease in interpreting the results, we report *Relative* MSE, namely the MSE divided by the MSE average when applying the kNN method. This shows that the MSE associated with kNN is roughly one half the MSE for KRAUS, a third of the MSE for $KL20$ and a fifth of the MSE for PACE. This may be observed from the boxplots of Relative MSE on the left side of Figure 1). On the bottom of this figure, we zoom in on these boxplots around the median by excluding the atypical values. The good results obtained by kNN are due in part to the sample size considered ($n = 2786$). However, by simulation we observe that, even for small values of n, kNN is competitive, becoming superior when n increases. Fig. A2 at the

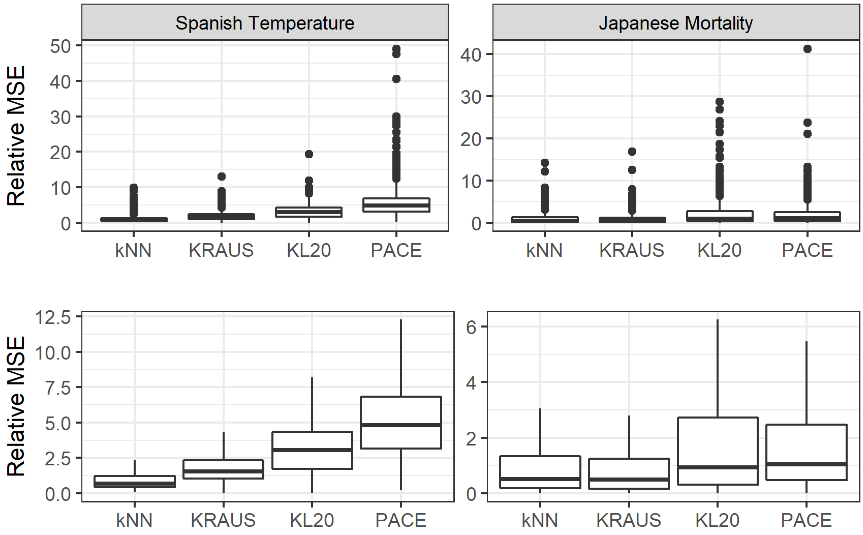

Fig. 1 Top panels: Boxplots of Relative MSE from 1000 reconstruction exercises based on yearly curves of Spanish daily temperatures and Japanese mortality rates. Bottom panels: Blown up images around the median of the above boxplots by excluding atypical values. The kNN estimator outperforms the competitors when reconstructing Spanish temperatures and it is competitive when reconstructing Japanese mortalities, where the curves are smoother and present local lineal patterns.

Appendix summarizes the results obtained by simulation. In addition, although all the methods are computationally efficient, kNN was the most efficient (see Table 1).

Figure 2 (left panels) illustrates the typical performance of each method. Although all the methods estimate correctly the mean temperature over the daily range, their estimated curve may be somewhat flattened, without the typical oscillations of Spanish daily temperatures. Only the kNN method provide estimators which may catch both the values of the curve and its shape. An additional attraction of kNN is its easy interpretation. The right panels of Figure 2 depict the kNN of the curve under reconstruction, showing that the curves used for reconstruction come from stations sharing similar weather in a roughly similar time period.

3.2.2 Japanese age-specific mortality rates

The Human Mortality Database (https://www.mortality.org) provides detailed mortality and population data of 41 countries or areas. For some countries, they also offer micro information by subdivision of the territory, providing data rich in spatio-temporal information. A FDA approach to analyze mortality data is to consider age-specific mortality rates as sample functions (Gao

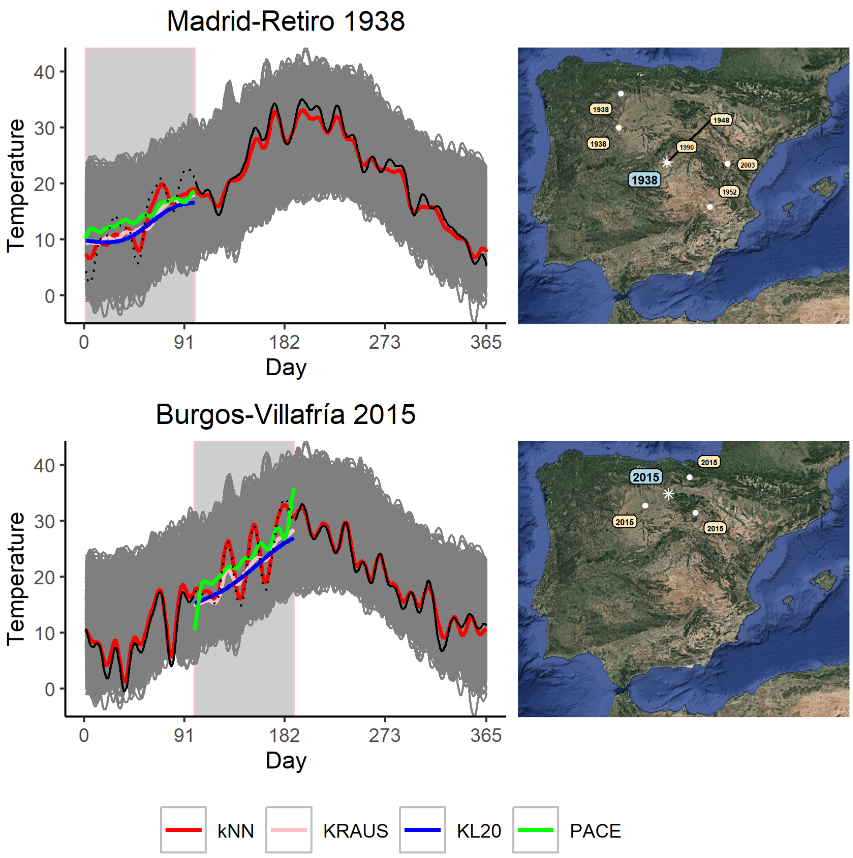

Fig. 2 Two illustrations of performance. The randomly observed part of the reconstructed curve is plotted as a black solid line whereas the censored part is dotted. We show both spatial and temporal location of the curves used for reconstruction by kNN. For reconstructing Madrid-Retiro 1938, the method chose $k = 2$, with Zamora 1938 being the 1st NN and Madrid-Retiro 1931 being the 2nd NN. For reconstructing Burgos-Villafría 2015, the method chose $k = 3$, with Palencia-Autilla Pino 2015 being the 1st NN, Foronda-Txokiza 2015 the 2nd NN, and Soria 2015 the 3rd NN. The chosen weather stations are very similar in climatological terms to the station with partially observed data.

et al., 2019; Shang and Hyndman, 2017). In particular, the Japanese mortality dataset is available for its 47 prefectures in many years. However, the curves of some prefectures are incomplete during some years. In total, we obtained 2007 complete curves of Japanese age-specific mortality rates with data between 1975 and 2016. We used these curves for comparing the reconstruction methods under consideration by repeating the simulation setup described in the previous subsection. All the methods perform well on these data. Typically they do not have the strong oscillations exhibited by the Spanish temperatures (see the Appendix A for some illustrations). In fact, there is a large subset of the domain where age-specific mortality curves present local lineal patterns (See Figure A3, ages between 25 and 100). KRAUS performed better than kNN, although the difference was negligible. Both methods worked better than PACE and KL20. These results are summarized on boxplots of MSE in Figure 1, as done already with the simulations based on Spanish temperatures. The computational efficiency of kNN is reported in Table 1.

4 Classification and outlier detection

After a brief introduction to the classification problem, we introduce a classification method based on localization distances, as well as on their Gaussian fluctuations, as given by Theorem 2.4. One novelty of the new approach is that it provides a probabilistic classifier, making it useful when combined with deterministic classifiers, such as the classifier based on global kNN. The other novelty is that localization distances may be used for outlier detection. Both tools, classifiers and outlier detection based on localization distances, are compared with other methods by using well known test data.

Assume that each curve $X_i, 1 \leq i \leq n$, comes from one of G groups (subpopulations). Let Y_i be the group label of X_i. That is to say Y_i equals y if X_i comes from group y, $1 \leq y \leq G$. Given the *training sample* $\{(X_i, Y_i)\}_{i=1}^n$ and a new curve X, the problem consists of predicting the label Y of X. An *ordinary* classifier is a rule that assigns to X a group label $m(X)$. Instead of outputting a group that X should belong to, a *probabilistic* classifier is a prediction of the conditional probability distribution of Y.

There exists a wide variety of methods for classifying functional data (Wang et al., 2016). Beyond treating the functional data as simple multivariate data in high dimensional spaces, many of these techniques make use of the fact that they are functions. For example this is done by adding their derivatives, integrals, and/or other preprocessing functions to the analysis (Hubert et al., 2017). The functional kNN classifier (fkNN) is a straightforward extension of the multivariate rule. In a nutshell, one considers the k nearest neighbors to the target curve and classifies it with the more represented group. This is the group to which the largest number of the k nearest neighbors belong. Then k is chosen to minimize the empirical misclassification rate on the training sample. This procedure is a benchmark method for classifying functional data but, as with other methods in the literature, it is deterministic. Next we introduce a probabilistic method. In particular, we provide a probability of right classification for fkNN.

4.1 The localization classifier

Let I_y be the set of indexes $i \in \{1, ..., n\}$ for which $Y_i = y$. Let $t_1, t_2, ..., t_M$ be the time points at which the data are observed. Although in practice they often come from a regular grid, in order to apply Theorem 2.4, we assume that they are i.i.d. uniform random variables on $[0, 1]$. Consider the empirical localization distance between X and the group y. This is

$$L_y(X) = \frac{1}{M} \sum_{r=1}^{M} L^{(k)}(X(t_r), \{X_j(t_r), j \in I_y\}). \tag{19}$$

Consider also its mean and variance

$$\mu_y = \frac{1}{M} \sum_{r=1}^{M} \mathbb{E}\big[L^{(k)}(X(t_r), \{X_j(t_r), j \in I_y\})\big] \quad \text{and} \quad \sigma_y^2 = \mathrm{Var}[L_y(X)]. \tag{20}$$

Finally, consider the standardized score $\tau(X, y) = (L_y(X) - \mu_y)/\sigma_y$. Denote by T the random variable $T = \tau(X, Y)$. We remark that T not only depends on (X, Y) but also on the training sample $\{(X_i, Y_i)\}_{i=1}^n$. Assume the conditional distribution of T given $\{Y = y\}$ is absolutely continuous with conditional probability density f_y and denote $\pi_y = \mathbb{P}(Y = y)$. Then the Bayes rule implies

$$\mathbb{P}(Y = y \mid T) = \frac{\pi_y f_y(T)}{\sum_{g=1}^{G} \pi_g f_g(T)}.$$

Following the basic idea for functional discriminant analysis for classification (Wang et al., 2016), we consider the Bayes classifier

$$\begin{aligned} m^{(k)}(X) &= \arg\max_y \mathbb{P}(Y = y \mid T) \\ &= \arg\max_y \pi_y f_y(T). \end{aligned} \tag{21}$$

Theorem 2.4 implies that the conditional distribution of T given $\{Y = y\}$ may be approximated by a standard normal distribution when M is large. Therefore, one might expect that the conditional probability density f_y should be approximated by a standard normal density, denoted here by ϕ. If this is the case, we may consider the following approximation of the Bayes classifier (21):

$$\begin{aligned} \tilde{m}^{(k)}(X) &= \arg\max_y \pi_y \phi(\tau(X, y)) \\ &= \arg\max_y \pi_y \phi(\frac{L_y(X) - \mu_y}{\sigma_y}). \end{aligned} \tag{22}$$

Here we require knowledge of μ_y and σ_y. These values may be estimated from the training sample as follows. First, consider the empirical localization distance between X_i and its group Y_i. According to (19), this is

$$L_i = \frac{1}{M}\sum_{r=1}^{M} L^{(k)}(X_i(t_r), \{X_j(t_r), j \in I_{Y_i}\}).$$

Next, consider the sample mean and variance of the empirical localization distances for each group. They are

$$\bar{L}_y = \frac{1}{n_y}\sum_{i=1}^{n} L_i \mathbf{1}_{\{Y_i=y\}} \quad \text{and} \quad S_y^2 = \frac{1}{n_y - 1}\sum_{i=1}^{n}(L_i - \bar{L}_y)^2 \mathbf{1}_{\{Y_i=y\}}. \tag{23}$$

These are empirical estimators of μ_y and σ_y^2 in (20). Thus, by plugging these estimators into (22), we obtain the empirical classifier

$$\eta^{(k)}(X) = \arg\max_{y} \pi_y \phi\big(\frac{L_y(X) - \bar{L}_y}{S_y}\big).$$

If M is large and n_y is also large for any group label y, we expect that $\eta^{(k)}(X)$ is similar to the Bayes classifier $m^{(k)}(X)$. Indeed, what we expect is

$$\mathbb{P}(Y = y \mid T) \approx \frac{\pi_y \phi((L_y(X) - \bar{L}_y)/S_y)}{\sum_{g=1}^{G} \pi_g \phi((L_g(X) - \bar{L}_g)/S_g)}.$$

We remark that, although the local feature of the empirical localization distances makes them robust in the presence of outliers, this is not the case of the sample mean and variance in (23). The accuracy of these estimators may be clearly affected by the presence of outlier data. For these reasons we consider trimmed means and variance, by discarding the outliers of each group, when calculating $\bar{L}_y$ and S_y^2 in (23). The problem of outlier detection in a group, say group y, is tackled by considering standard boxplots of the samples of empirical localization distances $\{L_i : Y_i = y\}$. This tool is simple but powerful given the asymptotic Gaussianity of the empirical localization distances. The problem of outlier detection in the complete population sample is addressed in a similar way by considering only one group. As far as we know, global fkNN has not been used for outlier detection, in fact we do not know a way to do it. Thus, the localization distances provide not only an estimate of classification for any deterministic rule, but they are also a useful tool for outlier detection.

If two or more groups are similar both in shape and scale, making the classification difficult, then different k values may provide different labels. The same occurs when one applies fkNN: different nearest neighbors may belong to different groups. In line with fkNN, we classify according to the more represented group. In our case, we use the modal label $\eta^{(1)}(X), \ldots, \eta^{(k)}(X)$. Also, as with f$k$NN, the k value is chosen to minimize the empirical misclassification rate on the training sample. We refer to this classification method by the *Localization classifier, or LC for short.*

To evaluate the proposed methodology we performed a comparative study with benchmark methods for both outlier detection and classification.

4.2 Classification study

We consider two types of benchmark functional classifiers. On the one hand, we consider functional extensions of the Depth-to-Depth classifier (DD) (Li et al., 2012) and on the other hand, we consider the classifiers introduced by Hubert et al. (2017), which also are inspired by DD but are based on special distances introduced by the cited authors instead of depths. Both approaches map functions to points on the plane which require classifying by some bivariate classifier such as kNN. The theory and methods for the former are discussed by Cuesta-Albertos et al. (2017) whereas Febrero-Bande and Oviedo (2012) developed the corresponding R package. The R package related to the classifiers introduced by Hubert et al. (2017) is also available at The Comprehensive R Archive Network (Segaert et al., 2019). Following these authors, we used kNN for bivariate classification; both kNN and fkNN were based on L^2-distances. For each methodology we considered the three classifiers suggested by their authors, corresponding to different choices of depth and distance. They are:

1. fAO, fBD and fSDO (Hubert et al., 2017).
2. DD_hM, DD_FM and DD_MBD (Cuesta-Albertos et al., 2017).

These methods are compared with fkNN and LC.

We tested the eight methods under consideration on the following three examples considered in the literature to which we have referred:

1. The fighter plane dataset used by Hubert et al. (2017). These are 210 univariate functions obtained from digital pictures of seven types of fighter planes, 30 from each type.
2. First derivative of the *Berkeley growth study*. This dataset contains the heights of 39 boys and 54 girls from ages 1 to 18 and is a classic in the literature of FDA (Ramsay and Silverman, 2005).
3. Second derivative of fat absorbance from the *Tecator* data set used by Cuesta-Albertos et al. (2017). For each piece of finely chopped meat we observe one spectrometric curve which corresponds to the absorbance measured at 100 wavelengths. The pieces are classified in Ferraty and Vieu (2006) into one of two classes according to small or large fat content. There are 12 pieces with low content and 103 with high.

Complete descriptions of the datasets appear in Cuesta-Albertos et al. (2017) and Hubert et al. (2017). From each group of these datasets, we randomly selected one half of the data for a training sample and we classified the rest. We repeated this process one thousand times and reported the missed classification rates on boxplots in Figure 3.

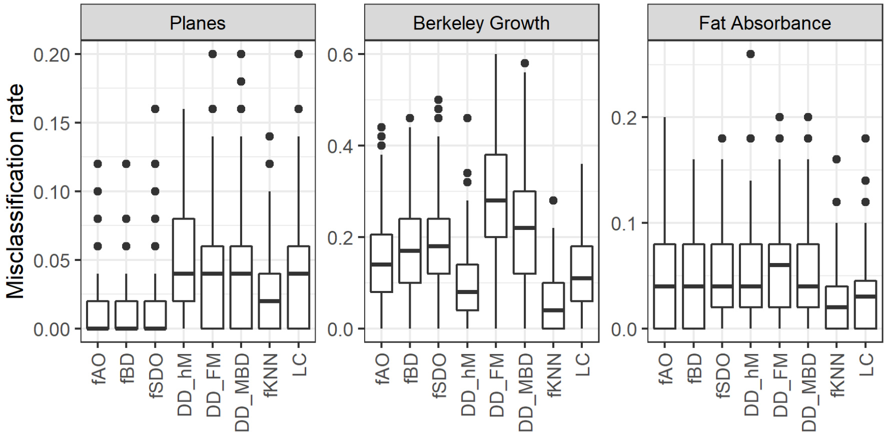

Fig. 3 Misclassification rates in 1000 runs of each considered dataset (fighter planes, first derivative of the Berkeley growth curves and second derivative of fat absorbance). For each classifying exercise, one half of the sample curves was randomly selected as a training sample.

In summary, although all the methods perform well for the fighter plane dataset, the methods of Hubert et al. (2017) were superior for these particular data. The methods yielded larger misclassifications rates for the other two datasets, where fkNN was the best option, closely followed by LC. Only DD_hM was competitive with LC but at the cost of a long computational running time. The pair (kNN, LC) offers a good classification tool, combining the efficient ordinary classification by fkNN with the predicted probability provided by LC.

4.3 Outlier detection study

One of the more popular tools for functional outlier detection is the functional boxplot (Sun and Genton, 2011). This method mimics the univariate boxplot by ordering the sample curves from the 'median' outward according to the modified band depth (López-Pintado and Romo, 2009). It is well known that the functional boxplot detects magnitude outliers, curves which are outlying in part the observation domain. However, the plot does not necessarily detect shape outliers. They are sample functions that have different shapes from the bulk of data. The outliergram (Arribas-Gil and Romo, 2014) and the MS-plot (Dai and Genton, 2018, 2019) were introduced for tackling both magnitude and shape outliers.

We compare the above three methods with the method based on localization processes. For this, we consider the Japanese mortality dataset discussed in Section 3. For each of the 47 prefectures we computed the average of the age-specific mortality rates between 1975 and 2007. We only used data until 2007 because data from the Saitama prefecture is not available after this date. In this way, we obtain 47 mortality curves smoothed by averaging over each prefecture. Using the default parameters suggested by the authors, all the methods detected an outlier at Okinawa, where residents have famously lived longer than anywhere else in the world. In addition, the outliergram also detected outliers at Fukui and Kochi. Indeed, we observe that the curves corresponding to these two prefectures exhibit strong oscillations for ages below fifty years. These oscillations are not seen in the greater part of the prefectures, which have smoother curves. Regarding localization distances, the boxplots corresponding to Okinawa and Aomori were very extreme outliers for all the considered k values. We remark that, although the rest of the methods did not detect an outlier at Aomori, this prefecture has experienced the highest mortality rates for many years. In fact, Aomori has been already considered an outlier by Japanese health officials (O'Donoghue, 2019). For several values of k, the localization distances corresponding to Fukui, Kochi and Tokyo fell above the default whiskers of the corresponding boxplots. Indeed, we observe that Tokyo is a deep datum for ages above thirty years but it has extreme low values of mortality rates for ages below 25. Finally, only for a few values of k, Nagano, Shiga and Kanagawa provide outliers with respect to localization processes and they fell close to the default whiskers. In fact, by considering the more conservative upper whisker ($\mathrm{Q}_3 + 3 * \mathrm{IQR}$ instead of $\mathrm{Q}_3 + 1.5 * \mathrm{IQR}$) neither Tokyo, Nagano, Shiga nor Kanagawa would be considered to be outliers. However, it is interesting to observe that both Nagano and Shiga show shapes similar to Fukui and Kochi, although with more moderate oscillations. Also, the behavior of Kanagawa is similar to Tokyo, but with less extreme mortality rates for ages below 25. The only value for which the eight mentioned prefectures fall above the default whiskers is $k = 9$. In Figure 4 we plot the outputs of all the methods under consideration whereas in Figure A3 of the Appendix A we show the particular case $k = 9$, where the reader can inspect the curve of each prefecture under consideration.

In conclusion, though localization distance statistics agreed with the three benchmark methods with respect to Okinawa and with the outliergram for Fukui and Kochi, they recognize Okinawa as an extreme outlier. In addition, the localization distances were able to detect Aomori as an extreme outlier and indicate a certain atypicality of Tokyo. Also, they drew attention to a small departure from the bulk of prefectures of Nagano, Shiga and Kanagawa.

5 Discussion

Localization processes provide an alternative to approximating curves from a given functional sample. These processes can be seen as piecewise approximations (of different orders) to a function from data collected in a functional setting. Other estimation methods, including for example those based on functional kNN and model-generated curves, often consider distances on function spaces for measuring nearness. Unlike these methods, we consider the localization width process formed by the point-by-point distances between the

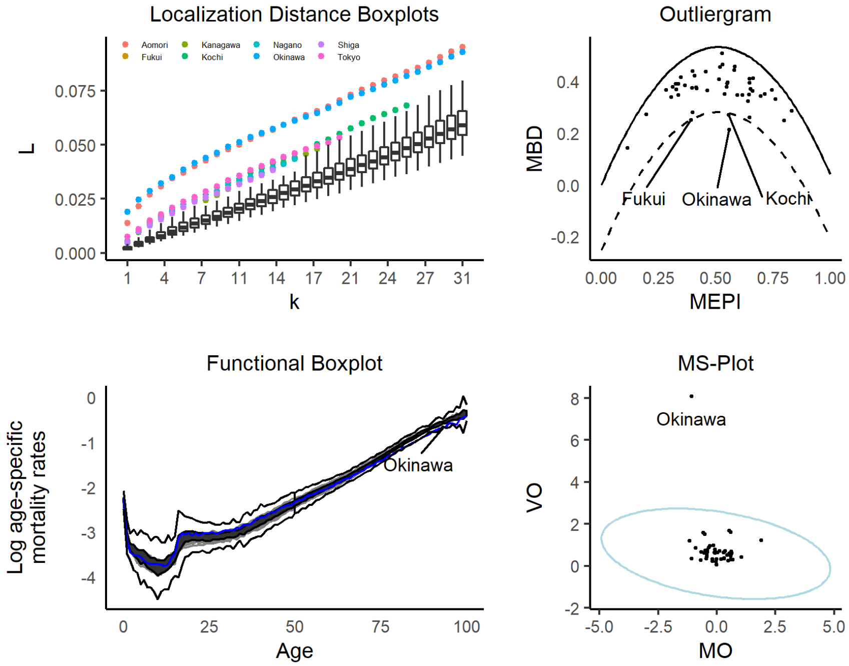

Fig. 4 Outputs from the outlier detection methods under consideration. All the methods detect Okinawa where residents have famously lived longer. Outliergram and localization distance boxplots agree with respect to Fukui and Kochi that exhibit strong oscillations for ages below fifty years. The localization distances are the only ones able to detect Aomori, a prefecture that experienced the highest mortality rates for many years. Also, localization distances are able to point out certain atypical features of Tokyo, Nagano, Shiga and Kanagawa. However these four prefectures are not extreme outliers, since their corresponding localization distances are smaller than conservative upper whiskers.

target curve and its corresponding approximation. Beyond the inherent interest of localization processes, we introduce them to provide a foundation for the rigorous asymptotic theory of nearest neighbor functional estimation.

First, we provide mean and distributional convergence of localization widths when the number of sample curves increases up to infinity. Under regularity conditions, we obtain $O(\frac{k}{2n})$ bounds on the expected L^1 norm of the difference between the kth localization process and the target. These results allow one to elucidate mild assumptions under which nearby neighbors to a target function on an observable range remain near the target outside this range. This property is the key to proving consistency of kNN type estimators for reconstructing curves from partially observed data. A particular kNN methodology is introduced and compared with three benchmark methods. We present results of a simulation study based on yearly curves of daily Spanish temperatures and Japanese age-specific mortality rates, two real world examples where a large range of contiguous data is missing, but which may be reconstructed. Beyond the intuitive appeal of the method, the results are promising in terms of accuracy, computational efficiency, and interpretability.

Second, the central limit theorem for empirical localization distances forms the basis of new methods for classification as well as outlier detection. These are problems where the kNN approach has been widely considered. A comparative study shows that the classification method proposed here is competitive with several benchmark methods. Only kNN gave consistently superior results. However, the new method predicts classification probabilities and provides standard normal scores that help validate the outputs rather than to only generate them. This makes the method a useful complementary tool capable of providing probabilistic support to the ordinary kNN classification. Regarding outlier detection, the case study considered shows that the method based on localization distances can detect both magnitude and shape outliers which go undetected by other methods.

In conclusion, the dual purpose of this paper has been to introduce the kNN localization processes, the associated kNN localization distances, as well as mathematical tools lending rigor to both the current approach and to possible further approximation schemes based on nearest neighbors. There remains the potential for further exploiting the asymptotic first and second order properties of localization distances in functional data analysis.

Supplementary information.

R-package localFDA: This provides routines for computing the localization processes and it performs classification and outlier detection.

Acknowledgments. The research of J. E. Yukich is supported in part by a Simons collaboration grant. He is also grateful for generous support from the Department of Statistics at Universidad Carlos III de Madrid, where most of this work was completed.

Declarations

- Funding: Not applicable.
- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use): Not applicable.
- Ethics approval: Not applicable.
- Consent to participate: Not applicable.
- Consent for publication: Not applicable.
- Availability of data and materials: the data is available from the sources indicated in the main text.
- Code availability: the R-package `localFDA` (Elías et al., 2021) provides routines for computing the localization processes and it performs classification and outlier detection.

Appendix A Proofs of main results and additional figures

Here we provide the proofs of the main results of Section 2, the proof of Proposition 1, and three additional figures.

A.1 Auxiliary results

We prepare for the proofs with two lemmas.

Lemma A.1. *Fix $k \in \mathbb{N}$. For almost all $t \in [0,1]$, there are random variables $\{X'_j(t)\}_{j=1}^n$, coupled to $\{X_j(t)\}_{j=1}^n$, and a Cox process $\mathcal{P}_{\kappa_t(X'_1(t))}$, also coupled to $\{X_j(t)\}_{j=1}^n$, such that as $n \to \infty$*

$$\begin{aligned} {W'}_n^{(k)}(t) &= W^{(k)}(X'_1(t), \{X'_j(t)\}_{j=1}^n) \\ &= \frac{2}{k} L^{(k)}(\mathbf{0}, n(\{X'_j(t)\}_{j=1}^n - X'_1(t))) \\ &\xrightarrow{\mathcal{P}} \frac{2}{k} L^{(k)}(\mathbf{0}, \mathcal{P}_{\kappa_t(X'_1(t))}) =: {W'}_\infty^{(k)}(t). \end{aligned}$$

Proof. The convergence may be deduced from Section 3 of Penrose and Yukich (2003) and we provide details as follows. For $x \in \mathbb{R}^d$ and $r > 0$ let $B(x,r)$ denote the Euclidean ball centered at x with radius r. Note that $L^{(k)}$ is a stabilizing score function on Poisson input $\mathcal{P}$ having constant intensity density, that is to say that its value at the origin is determined by the local data consisting of the points in the intersection of the realization of $\mathcal{P}$ and the ball $B(\mathbf{0}, R^{L^{(k)}}(\mathbf{0},\mathcal{P}))$, where $R^{L^{(k)}}(\mathbf{0},\mathcal{P})$ is a radius of stabilization. For precise definitions we refer to Penrose and Yukich (2003), Section 3 and Penrose (2007).

The coupling of Section 3 of Penrose and Yukich (2003) shows that we may find $\{X'_j(t)\}_{j=2}^n$, where $X'_j(t) \overset{\mathcal{D}}{=} X_j(t), j = 2, ..., n$ and a Cox process $\mathcal{P}_{\kappa_t(X'_1(t))}$ such that if we put $\mathcal{X}'_{n-1}(t) = \{X'_j(t)\}_{j=2}^n$, then for all $K > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(n(\mathcal{X}'_{n-1}(t) - X'_1(t)) \cap B(\mathbf{0},K) = \mathcal{P}_{\kappa_t(X'_1(t))} \cap B(\mathbf{0},K)\right) = 1, \quad \text{(A1)}$$

which is a consequence of the convergence of the point process $n(\mathcal{X}'_{n-1}(t) - X'_1(t)) \cap B(\mathbf{0},K)$ to the point process $\mathcal{P}_{\kappa_t(X'_1(t))} \cap B(\mathbf{0},K)$ as $n \to \infty$. See Lemma 3.1 of Penrose and Yukich (2003) for details. Fix $\epsilon > 0$. Now write for all $\delta > 0$

$$\begin{aligned} &\mathbb{P}\left(|\frac{2}{k} L^{(k)}(\mathbf{0}, n(\mathcal{X}'_{n-1}(t) - X'_1(t))) - \frac{2}{k} L^{(k)}(\mathbf{0}, \mathcal{P}_{\kappa_t(X'_1(t))})| > \epsilon\right) \\ &\leq \mathbb{P}(n(\mathcal{X}'_{n-1}(t) - X'_1(t)) \cap B(\mathbf{0},K) \neq \mathcal{P}_{\kappa_t(X'_1(t))} \cap B(\mathbf{0},K)) \\ &\qquad + \mathbb{P}(R^{L^{(k)}}(\mathbf{0}, \mathcal{P}_{\kappa_t(X'_1(t))}) > K) \end{aligned}$$

$$\leq \mathbb{P}(n(\mathcal{X}'_{n-1}(t) - X'_1(t)) \cap B(\mathbf{0}, K) \neq \mathcal{P}_{\kappa_t(X'_1(t))} \cap B(\mathbf{0}, K)) \\ + \mathbb{P}(R^{L^{(k)}}(\mathbf{0}, \mathcal{P}_{\kappa_t(X'_1(t))}) > K, \kappa_t(X'_1(t)) \geq \delta) + \mathbb{P}(\kappa_t(X'_1(t)) \leq \delta). \tag{A2}$$

Given $\epsilon > 0$, the last term in (A2) may be made less than $\epsilon/3$ if δ is small. The penultimate term is bounded by $\mathbb{P}(R^{L^{(k)}}(\mathbf{0}, \mathcal{P}_\delta) > K)$, which is less than $\epsilon/3$ if K is large, since $R^{L^{(k)}}(\mathbf{0}, \mathcal{P}_\delta)$ is finite a.s. By (A1) the first term is less than $\epsilon/3$ if n is large. Thus, for δ small and K and n large, the right-hand side of (A2) is less than ϵ, which concludes the proof. □

Lemma A.2. *Assume that the data are bounded and regular from below for all $t \in T_0 \subseteq [0,1]$ as at (6), and where T_0 has Lebesgue measure 1. Then $\sup_{n\leq\infty} \sup_{t\in T_0} \mathbb{E}W_n^{(k)}(t)^2 \leq C$, where C is a finite constant.*

Proof. We treat the case $1 \leq n < \infty$, as the case $n = \infty$ follows by similar methods. Without loss of generality we may assume the data are bounded above by 1 and that $S(\kappa_t) = [0,1]$ for all $t \in T_0$. We first prove the lemma for $k = 1$ and then for general k. Let S_δ be the subinterval of $[0,1]$ such that $\kappa_t(x) \geq \kappa_{\min}$ for all $x \in S_\delta$. Note that $|S_\delta|= \delta \in (0,1]$ by assumption. We have for all $r > 0$ and all $t \in T_0$

$$\begin{aligned}\mathbb{P}(W_n^{(1)}(X_1(t), \{X_j(t)\}_{j=2}^n) \geq r) &= \mathbb{P}\left(L^{(1)}(X_1(t), \{X_j(t)\}_{j=2}^n) \geq \frac{r}{2n}\right) \\ &= \Pi_{j=2}^n \mathbb{P}\left(|X_1(t) - X_j(t)| \geq \frac{r}{2n}\right) \\ &= \left(1 - \mathbb{P}\left(|X_1(t) - X_2(t)| \leq \frac{r}{2n}\right)\right)^{n-1} \\ &= \left(1 - \int_{[0,1]} \int_{|x_1-x_2|\leq \frac{r}{2n}} \kappa_t(x_2)dx_2\kappa_t(x_1)dx_1\right)^{n-1}.\end{aligned}$$

By the boundedness assumption on the data, we have $\mathbb{P}(L^{(1)}(X_1(t), \{X_j(t)\}_{j=2}^n) \geq \frac{r}{2n}) = 0$ for all $r \in (2n, \infty)$. Thus we may assume without loss of generality that $r \in [0, 2n]$. We bound the double integral from below by

$$\begin{aligned}\int_{[0,1]} \int_{|x_1-x_2|\leq \frac{r}{2n}} \kappa_t(x_2)dx_2\kappa_t(x_1)dx_1 &\geq \int_{S_\delta} \int_{|x_1-x_2|\leq \frac{r}{2n}} \kappa_t(x_2)dx_2\kappa_t(x_1)dx_1 \\ &\geq \int_{S_\delta} \int_{|x_1-x_2|\leq \frac{r\delta}{2n}} \kappa_t(x_2)dx_2\kappa_t(x_1)dx_1 \\ &\geq c\delta\kappa_{\min}^2 \cdot \frac{r\delta}{2n},\end{aligned}$$

where here and elsewhere $c > 0$ is a generic constant, possibly changing from line to line. This gives

$$\mathbb{P}(W_n^{(1)}(X_1(t), \{X_j(t)\}_{j=1}^n) \geq r) \leq \left(1 - \delta\kappa_{\min}^2 \cdot \frac{r\delta}{2n}\right)^{n-1}$$
$$\leq c \exp(-\frac{\delta^2\kappa_{\min}^2 r}{c}),\ r > 0.$$

Random variables having exponentially decaying tails have finite moments of all orders and this proves the lemma for $k = 1$.

The proof for general k is similar. We show this holds for $k = 2$ as follows. We have

$$\mathbb{P}(W_n^{(2)}(X_1(t), \{X_j(t)\}_{j=2}^n) \geq r) = \mathbb{P}\left(L^{(2)}(X_1(t), \{X_j(t)\}_{j=2}^n) \geq \frac{r}{n}\right).$$

Given the event $\{L^{(2)}(X_1(t), \{X_j(t)\}_{j=2}^n) \geq \frac{r}{n}\}$, either the first nearest neighbor to $X_1(t)$ is at a distance greater than $\frac{r}{n}$ to $X_1(t)$ or the first nearest neighbor to $X_1(t)$ is at a distance less than $\frac{r}{n}$ to $X_1(t)$ and the first nearest neighbor among the remaining $n-2$ sample points exceeds $\frac{r}{n}$.

This gives

$$\mathbb{P}\left(L^{(2)}(X_1(t), \{X_j(t)\}_{j=2}^n) \geq \frac{r}{n}\right) \leq \mathbb{P}\left(L^{(1)}(X_1(t), \{X_j(t)\}_{j=2}^n) \geq \frac{r}{n}\right)$$
$$+ \sum_{i=2}^n \mathbb{P}\left(L^{(1)}(X_1(t), \{X_j(t)\}_{j=2, j\neq i}^n) \geq \frac{r}{n}\right) \mathbb{P}(|X_1 - X_i| \leq \frac{r}{n}).$$

Since $\mathbb{P}(|X_1 - X_i| \leq \frac{r}{n}) = O(n^{-1})$ for all $i = 2, ..., n$ we may use the bounds for the case $k = 1$ to show that $\mathbb{P}\left(L^{(2)}(X_1(t), \{X_j(t)\}_{j=2}^n) \geq \frac{r}{n}\right)$ decays exponentially fast in r. The proof for general k follows in a similar fashion and we leave the details to the reader. This proves the lemma. □

A.2 Proof of Theorem 2.1

We first prove (3). Let $t \in [0, 1]$ be such that the marginal density κ_t exists. By translation invariance of $L^{(k)}$ we have as $n \to \infty$

$$\mathbb{E}W_n^{(k)}(t) = \mathbb{E}L^{(k)}\left(\frac{2n}{k}X_1(t), \frac{2n}{k}\{X_j(t)\}_{j=1}^n\right)$$
$$= \frac{2}{k}\mathbb{E}L^{(k)}(\mathbf{0}, n(\{X_j(t)\}_{j=1}^n - X_1(t)))$$
$$\to \int_{S(\kappa_t)} \frac{2}{k}\mathbb{E}L^{(k)}(\mathbf{0}, \mathcal{P}_{\kappa_t(y)})\kappa_t(y)dy,$$

where the limit follows since convergence in probability (Lemma A.1) combined with uniform integrability (Lemma A.2) gives convergence in mean. For any constant τ we have $\mathbb{E}L^{(k)}(\mathbf{0}, \mathcal{P}_\tau) = \tau^{-1}\mathbb{E}L^{(k)}(\mathbf{0}, \mathcal{P}_1)$. Notice that $L^{(k)}(\mathbf{0}, \mathcal{P}_1)$ is a Gamma $\Gamma(k, 2)$ random variable with shape parameter k and scale parameter 2 and thus $\mathbb{E}L^{(k)}(\mathbf{0}, \mathcal{P}_1) = k/2$. The proof of (3) is complete.

To prove (4), we replace $W_n^{(k)}(t)$ by its square in the above computation. This yields

$$\lim_{n\to\infty} \mathbb{E}W_n^{(k)}(t)^2 = \int_{S(\kappa_t)} \frac{4}{k^2}\mathbb{E}(L^{(k)}(\mathbf{0}, \mathcal{P}_{\kappa_t(y)}))^2\kappa_t(y)dy.$$

For any constant $\tau \in (0, \infty)$ we have

$$\mathbb{E}(L^{(k)}(\mathbf{0}, \mathcal{P}_\tau))^2 = \tau^{-2}\mathbb{E}(L^{(k)}(\mathbf{0}, \mathcal{P}_1))^2 = \tau^{-2}\frac{(k+1)k}{4}, \tag{A3}$$

where we recall that the second moment of a Gamma $\Gamma(k, 2)$ random variable equals $(k + 1)k/4$. These facts yield (4). The limit (5) is a consequence of Lemma A.1. □

A.3 Proof of Theorem 2.2

Recall that on $T_0 \subseteq [0, 1]$ we have that κ_t exists, the data are bounded, and the data are regular from below, as at (6). By Lemmas A.1 and A.2, the random variables $W_n'^{(k)}(t) = W^{(k)}(X_1'(t), \{X_j'(t)\}_{j=1}^n), t \in T_0$, converge in probability and also in mean. It follows that for all $t \in T_0$ as $n \to \infty$

$$F_n(t) = \mathbb{E}|W_n'^{(k)}(t) - W_\infty'^{(k)}(t)| = \mathbb{E}|W_n^{(k)}(t) - W_\infty^{(k)}(t)| \to 0.$$

Now $\sup_n \sup_{t\in T_0} F_n(t) \le C$ and the bounded convergence theorem gives $\lim_{n\to\infty} \int_0^1 F_n(t)dt = 0$. This gives the first statement of Theorem 2.2. The identity $\lim_{n\to\infty} \int_0^1 \mathbb{E}W_\infty^{(k)}(t)dt = 1$ follows from

$$\int_0^1 F_n(t)dt \ge |\int_0^1 \mathbb{E}W_n^{(k)}(t)dt - \int_0^1 \mathbb{E}W_\infty^{(k)}(t)dt|$$

and the identity $\mathbb{E}W_\infty^{(k)}(t) = |S(\kappa_t)|$, $t \in T_0$. □

A.4 Proof of Theorem 2.3

Lemma A.1 assumes that k is fixed. The lemma will not always hold if k is growing arbitrarily fast with n. Thus our proof techniques and coupling arguments need to be modified. We break the proof of Theorem 2.3 into five parts.

Part (i) Coupling. We start with a general coupling fact. Given Poisson point processes Σ_1 and Σ_2 with densities f_1 and f_2, we may find coupled Poisson point processes Σ_1' and Σ_2' with $\Sigma_1' \overset{\mathcal{D}}{=} \Sigma_1$ and $\Sigma_2' \overset{\mathcal{D}}{=} \Sigma_2$ such that the probability that the two point processes are not equal on $[-A, A]$ is bounded by

$$\int_{-A}^{A} |f_1(x) - f_2(x)| dx.$$

Let $t \in [0,1]$ be such that the marginal density κ_t exists. As in Theorem 2.3, we assume that κ_t is α-Hölder continuous for $\alpha \in (0,1]$. Note that the point process $n(\mathcal{P}_{n\kappa_t} - y)$ has intensity density $\kappa_t(\frac{x}{n} + y)$, $x \in n(S(\kappa_t) - y)$. For each $y \in S(\kappa_t)$, we may find coupled Poisson point processes $\mathcal{P}'_{n\kappa_t}$ and $\mathcal{P}'_{\kappa_t(y)}$ with $n(\mathcal{P}'_{n\kappa_t} - y) \overset{\mathcal{D}}{=} n(\mathcal{P}_{n\kappa_t} - y)$ and $\mathcal{P}'_{\kappa_t(y)} \overset{\mathcal{D}}{=} \mathcal{P}_{\kappa_t(y)}$ such that the probability that the point processes $n(\mathcal{P}'_{n\kappa_t} - y)$ and $\mathcal{P}'_{\kappa_t(y)}$ are not equal on $[-A, A]$ is bounded uniformly in $y \in S(\kappa_t)$ by

$$\int_{-A}^{A} |\kappa_t\left(\frac{x}{n} + y\right) - \kappa_t(y)| dx \leq 2A \left(\frac{A}{n}\right)^{\alpha}. \tag{A4}$$

We will need this coupling in what follows.

Part (ii) Poissonization. We will first show a Poissonized version of (9). Write k instead of $k(n)$. We assume that we are given a Poisson number of data curves $\{X_j(t)\}_{j=1}^{N(n)}$, where $N(n)$ is an independent Poisson random variable with parameter n. We aim to show for almost all $t \in [0,1]$

$$\lim_{n\to\infty} \mathbb{E} W_n^{(k)}(X_1(t), \{X_j(t)\}_{j=1}^{N(n)}) = |S(\kappa_t)|. \tag{A5}$$

By translation invariance of $W_n^{(k)}$ we have

$$\begin{aligned} \mathbb{E} W_n^{(k)}(X_1(t), \{X_j(t)\}_{j=1}^{N(n)}) &= \frac{2n}{k} \mathbb{E} L^{(k)}\left(X_1(t), \{X_j(t)\}_{j=1}^{N(n)}\right) \\ &= \frac{2n}{k} \mathbb{E} L^{(k)}(\mathbf{0}, (\{X_j(t)\}_{j=1}^{N(n)} - X_1(t))). \end{aligned} \tag{A6}$$

We assert that as $n \to \infty$

$$|\mathbb{E} L^{(k)}(\mathbf{0}, \frac{2n}{k}(\{X_j(t)\}_{j=1}^{N(n)} - X_1(t))) - \frac{2}{k} \mathbb{E} L^{(k)}(\mathbf{0}, \mathcal{P}_{\kappa_t(X_1(t))})| \to 0. \tag{A7}$$

Combining (A6)-(A7) and recalling that $\frac{2}{k}\mathbb{E} L^{(k)}(\mathbf{0}, \mathcal{P}_{\kappa_t(X_1(y))}) = |S(\kappa_t)|$, we obtain

$$\lim_{n\to\infty} \mathbb{E} W_n^{(k)}(X_1(t), \{X_j(t)\}_{j=1}^{N(n)}) = |S(\kappa_t)|,$$

which establishes (A5).

It remains to establish (A7). A Poisson point process on a set S with intensity density $nh(x)$, where h is itself a density, may be expressed as the realization of random variables $X_1, ..., X_{N(n)}$, where $N(n)$ is an independent Poisson random variable with parameter n and where each X_i has density h on S. Thus the point process $\{X_j(t)\}_{j=1}^{N(n)}$ is the Poisson point process $\mathcal{P}_{n\kappa_t}$. To show the assertion (A7) we thus need to show

$$|\frac{2}{k}\mathbb{E}L^{(k)}(\mathbf{0}, n(\mathcal{P}_{n\kappa_t} - X_1(t))) - \frac{2}{k}\mathbb{E}L^{(k)}(\mathbf{0}, \mathcal{P}_{\kappa_t(X_1(t))})| \to 0$$

or equivalently,

$$|\frac{1}{k}\mathbb{E}L^{(k)}(\mathbf{0}, n(\mathcal{P}'_{n\kappa_t} - X'_1(t))) - \frac{1}{k}\mathbb{E}L^{(k)}(\mathbf{0}, \mathcal{P}'_{\kappa_t(X'_1(t))})| \to 0,$$

where $\mathcal{P}'_{n\kappa_t}$ and $\mathcal{P}'_{\kappa_t(y)}$ are as in part (i).

Fix $\epsilon > 0$. As in the bound (A2), we have for all $\delta > 0, K > 0$

$$\begin{aligned}
&\mathbb{P}\left(|\frac{2}{k}L^{(k)}(\mathbf{0}, n(\mathcal{P}'_{n\kappa_t} - X'_1(t))) - \frac{2}{k}L^{(k)}(\mathbf{0}, \mathcal{P}'_{\kappa_t(X'_1(t))})| > \epsilon\right)\\
&\le \mathbb{P}\left(\frac{2n}{k}(\mathcal{P}'_{n\kappa_t} - X'_1(t)) \cap B(\mathbf{0}, K) \neq \frac{2}{k}\mathcal{P}'_{\kappa_t(X'_1(t))} \cap B(\mathbf{0}, K)\right)\\
&\qquad + \mathbb{P}\left(R^{L^{(k)}}(\mathbf{0}, \frac{2}{k}\mathcal{P}'_{\kappa_t(X'_1(t))}) > K\right)\\
&\le \mathbb{P}\left(\frac{2n}{k}(\mathcal{P}'_{n\kappa_t} - X'_1(t)) \cap B(\mathbf{0}, K) \neq \frac{2}{k}\mathcal{P}'_{\kappa_t(X'_1(t))} \cap B(\mathbf{0}, K)\right)\\
&\qquad + \mathbb{P}\left(R^{L^{(k)}}(\mathbf{0}, \frac{2}{k}\mathcal{P}'_{\kappa_t(X'_1(t))}) > K, \kappa_t(X'_1(t)) \ge \delta\right) + \mathbb{P}(\kappa_t(X'_1(t)) \le \delta).
\end{aligned} \tag{A8}$$

The last term in (A8) may be made less than $\epsilon/3$ if δ is small. The penultimate term is bounded $\mathbb{P}(R^{L^{(k)}}(\mathbf{0}, \mathcal{P}_{\delta k}) > K) = \mathbb{P}(R^{L^{(k)}}(\mathbf{0}, \frac{1}{\delta k}\mathcal{P}_1) > K) = \mathbb{P}(\frac{1}{\delta k}\Gamma(k, 2) > K)$, which by Chebyshev's inequality is less than $\epsilon/3$ if K is large. The first term satisfies

$$\begin{aligned}
&\mathbb{P}\left(\frac{2n}{k}(\mathcal{P}'_{n\kappa_t} - X'_1(t)) \cap B(\mathbf{0}, K) \neq \frac{2}{k}\mathcal{P}'_{\kappa_t(X'_1(t))} \cap B(\mathbf{0}, K)\right)\\
&= \mathbb{P}\left(n(\mathcal{P}'_{n\kappa_t} - X'_1(t)) \cap B(\mathbf{0}, \frac{Kk}{2}) \neq \mathcal{P}'_{\kappa_t(X'_1(t))} \cap B(\mathbf{0}, \frac{Kk}{2})\right)\\
&\le Kk(\frac{Kk}{2n})^{\alpha}
\end{aligned} \tag{A9}$$

where the inequality follows from (A4). By assumption, we have $\lim_{n\to\infty} \frac{k^{1+\alpha}}{n^{\alpha}} = 0$ and it follows that the first term is less than $\epsilon/3$ if n is large.

Thus, for δ small and K and n large, the right-hand side of (A2) is less than ϵ. Thus

$$|\frac{2}{k}L^{(k)}(\mathbf{0}, n(\mathcal{P}'_{\kappa_t} - X'_1(t))) - \frac{2}{k}L^{(k)}(\mathbf{0}, \mathcal{P}'_{\kappa_t(X'_1(t))})| \xrightarrow{\mathcal{P}} 0.$$

The assertion (A7) follows since convergence in probability combined with uniform integrability gives convergence in mean.

Part (iii) de-Poissonization. We de-Poissonize the above equality to obtain (9). In other words we need to show that the limit does not change when $N(n)$ is replaced by n. Put

$$\mathcal{Y}_n = \begin{cases} X_1(t), ..., X_{N(n)-(N(n)-n)^+}(t), & \text{if } N(n) \geq n \\ X_1(t), ..., X_{N(n)+(n-N(n))^+}(t), & \text{if } N(n) < n. \end{cases}$$

Then $\mathcal{Y}_n \overset{\mathcal{D}}{=} \{X_1(t), X_2(t), ..., X_n(t)\}$. We use this coupling of Poisson and binomial input in all that follows.

We wish to show that $\hat{X}^{(k)}_{n,1}(t)$ coincides with $\hat{X}^{(k)}_{N(n),1}(t)$ on a high probability event; in other words we wish to show that the sample points with indices between $\min(n, N(n))$ and $\max(n, N(n))$ do not, in general, modify the value of $\hat{X}^{(k)}_{n,1}(t)$. Consider the event that the Poisson random variable does not differ too much from its mean, i.e.,

$$E_n = \{|N(n) - n| \leq c\sqrt{n}\log n\}$$

and note that tail bounds for Poisson random variables show that there is $c > 0$ such that $\mathbb{P}(E_n^c) = O(n^{-2})$. Write

$$\begin{aligned} &\mathbb{E}|W_n^{(k)}(X_1(t), \{X_j(t)\}_{j=1}^{N(n)}) - W_n(t)| \\ &\leq \mathbb{E}|[W_n^{(k)}(X_1(t), \{X_j(t)\}_{j=1}^{N(n)}) - W_n^{(k)}(t)]\mathbf{1}(E_n)| \\ &\qquad + \mathbb{E}|[W_n^{(k)}(X_1(t), \{X_j(t)\}_{j=1}^{N(n)}) - W_n^{(k)}(t)]\mathbf{1}(E_n^c)|. \end{aligned}$$

The last summand is $o(1)$, which may be seen using the Cauchy-Schwarz inequality, Lemma A.2, and $\mathbb{P}(E_n^c) = O(n^{-2})$.

For any $1 \leq j \leq c\sqrt{n}\log n$ we define

$$A_{n,j} = A_{n,j}(t) = \{|X_{\min(n,N(n))+j}(t) - X_1(t)| \geq |\hat{X}^{(k)}_{n,1}(t) - X_1(t)|\}.$$

This is the event that the data curves having index larger than $\min(n, N(n))$ are farther away from $X_1(t)$ than is the data curve $\hat{X}^{(k)}_{n,1}(t)$.

Given i.i.d. random variables $Z_i, 1 \leq i \leq n$, we let $Z_i^{(k)}$ denote the kth nearest neighbor to Z_i. Given an independent random variable Z_0 having the same distribution as Z_i, the probability that Z_0 belongs to $[Z_i, Z_i^{(k)}]$ coincides with the probability that a uniform random variable on $[0,1]$ belongs to $[U_i, U_i^{(k)}]$ where $U_i, 1 \le i \le n$, are i.i.d. uniform random variables on $[0,1]$. By exchangeability this last probability equals $k/(n-1)$.

It follows that for any $j = 1, 2, \ldots$

$$\mathbb{P}(A_{n,j} \mid n \le N(n)) = \frac{(n-1)-k}{n-1},$$

whereas

$$\mathbb{P}(A_{n,j} \mid n \ge N(n)) = \frac{(N(n)-1)-k}{N(n)-1}.$$

We have

$$\mathbb{P}(A_{n,j} \mid N(n)) = 1 - \frac{k}{\min((n-1),(N(n)-1))}$$

and thus $\mathbb{P}(A_{n,j} \cap E_n) \ge 1 - \frac{k}{(n-1)-c\sqrt{n}\log n}$. Thus

$$\begin{aligned}
&\mathbb{P}((W_n^{(k)}(X_1(t), \{X_j(t)\}_{j=1}^{N(n)}) - W_n^{(k)}(t))\mathbf{1}_{\{E_n\}} \ne 0) \\
&= 1 - (\mathbb{P}(A_{n,1} \cap E_n))^{c\sqrt{n}\log n} \\
&\le 1 - \left(1 - \frac{k}{(n-1)-c\sqrt{n}\log n}\right)^{c\sqrt{n}\log n} \\
&\le \frac{kc'\log n}{\sqrt{n}}.
\end{aligned}$$

When $\frac{kc'\log n}{\sqrt{n}} = o(1)$ we find that $\left(W_n^{(k)}(X_1(t), \{X_j(t)\}_{j=1}^{N(n)}) - W_n^{(k)}(t)\right)\mathbf{1}(E_n)$ converges to zero in probability as $n \to \infty$, and thus so does $\left(W_n^{(k)}(X_1(t), \{X_j(t)\}_{j=1}^{N(n)}) - W_n^{(k)}(t)\right)$. By uniform integrability we obtain that $(W_n^{(k)}(X_1(t), \{X_j(t)\}_{j=1}^{N(n)}) - W_n^{(k)}(t))$ converges to zero in mean. This completes the proof of (9).

Part (iv) Variance convergence. Replacing $W_n^{(k)}(X_1(t), \{X_j(t)\}_{j=1}^{N(n)})$ by its square in the above computation gives

$$\begin{aligned}
&\lim_{n\to\infty} [\mathbb{E}W_n^{(k)}(X_1(t), \{X_j(t)\}_{j=1}^{N(n)})^2 - (\mathbb{E}W_n^{(k)}X_1(t), \{X_j(t)\}_{j=1}^{N(n)})^2] \\
&= \lim_{n\to\infty} \frac{4}{k^2} \int_{S(\kappa_t)} \mathbb{E}L^{(k)}(\mathbf{0}, \mathcal{P}_{\kappa_t(y)})^2 \kappa_t(y)dy - |S(\kappa_t)|^2 \\
&= \int_{S(\kappa_t)} \frac{1}{\kappa_t(y)} dy - |S(\kappa_t)|^2,
\end{aligned}$$

where the last equality makes use of (A3). This gives (10), as desired.

Part (v) L^1 convergence. The limit (11) follows exactly as in the proof of Theorem 2.2. □

A.5 Proof of Theorem 2.4

This result is a straightforward consequence of the central limit theorem for M-dependent random variables. It is enough to prove the central limit theorem for the re-scaled random variables $\{L(mT_r)\}_{r=1}^{m}$. Indeed these random variables $\{L(mT_r)\}_{r=1}^{m}$ have moments of all orders and they are M-dependent since $\{L(mT_r)\}_{r\in A}$ and $\{L(mT_r)\}_{r\in B}$ are independent whenever the distance between the index sets A and B exceeds $2M$. The asserted asymptotic normality follows by the classical central limit theorem for M-dependent random variables. □

A.6 Proof of Proposition 1

Proof. It is enough to show for any fixed j and all $\varepsilon > 0$ that

$$\lim_{n\to\infty} \mathbb{P}\Big(\int_0^1 |X^{(n,j)}(t) - X_1(t)|dt < \varepsilon\Big) = 1.$$

We have

$$\begin{aligned}
&\mathbb{P}\Big(\int_0^1 |X^{(n,j)}(t) - X_1(t)|dt > \varepsilon\Big)\\
&\le \mathbb{P}\Big(\int_0^1 |X^{(n,j)}(t) - X_1(t)|\mathbf{1}_{I^{(k)}(X^{(n,j)})}(t)dt > \frac{\varepsilon}{2}\Big)\\
&\quad + \mathbb{P}\Big(\int_0^1 |X^{(n,j)}(t) - X_1(t)|\mathbf{1}_{[0,1]\setminus I^{(k)}(X^{(n,j)})}(t)dt > \frac{\varepsilon}{2}\Big)\\
&\le \frac{2}{\varepsilon}\mathbb{E}\int_0^1 |\hat{X}_1^{(k)}(t) - X_1(t)|dt + \mathbb{E}\int_0^1 \mathbf{1}_{[0,1]\setminus I^{(k)}(X^{(n,j)})}(t)dt\\
&= \frac{k}{n\varepsilon}\mathbb{E}\int_0^1 W_n^{(k)}(t)dt + \mathbb{E}\int_0^1 \mathbf{1}_{[0,1]\setminus I^{(k)}(X^{(n,j)})}(t)dt. \qquad \text{(A10)}
\end{aligned}$$

Since κ_t is α-Hölder continuous with $\alpha = 1$, we may apply Theorem 2.3 for $\alpha = 1$ and $k = k(n) = o(\sqrt{n})$. Thus, as $n \to \infty$, the right-hand side goes to 0 by Theorem 2.3, the finiteness of $\mathbb{E}\int_0^1 W_n^{(k)}(t)dt$, and (18). □

A.7 Additional figures

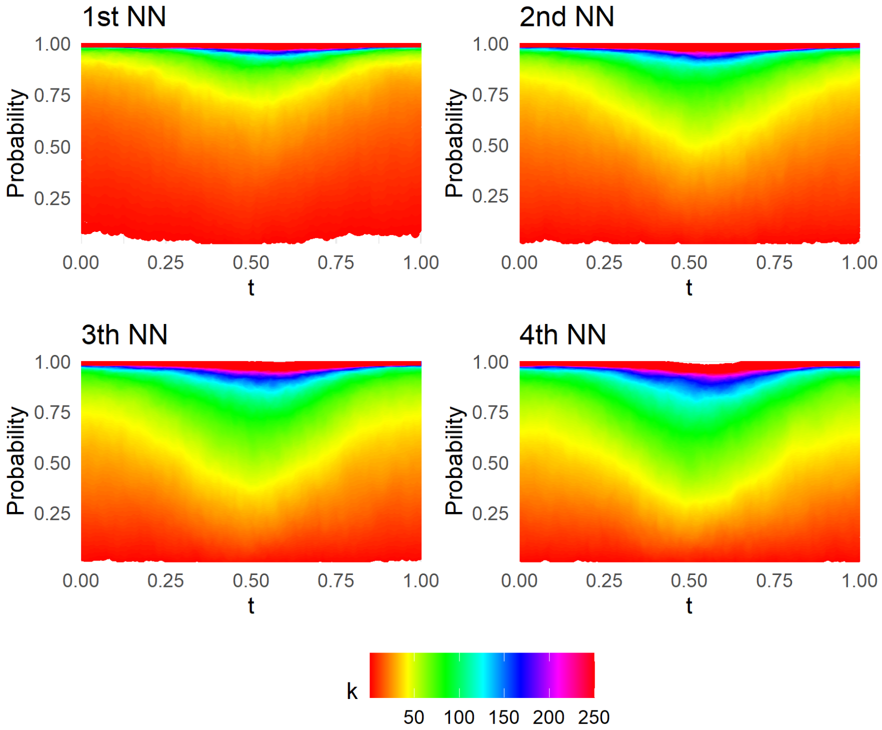

Fig. A1 Estimated values of $\mathbb{P}\left(I^{(k)}\left(X^{(n,j)}\right) \text{ contains } t\right)$, $0 \leq t \leq 1$, $1 \leq j \leq 4$, $1 \leq k \leq 250$, $n = 2500$. The estimation is based on 1000 independent replicates of (X_1, O_1) when O_1 is obtaining by removing randomly one interval of the partition of $[0, 1]$ induced by two independent Uniform(0,1) variables. X_1 is a linear combination of sines and cosines with independent Gaussian coefficients.

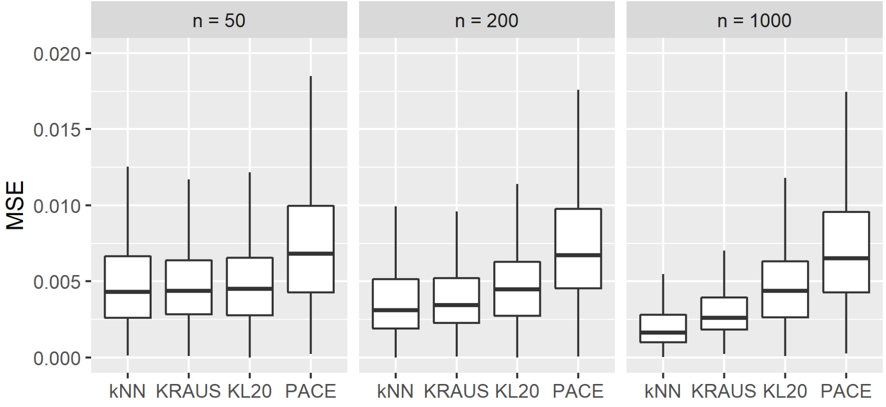

Fig. A2 Boxplots of Relative MSE from 1000 reconstruction exercises based on 50, 200 and 1000 curves randomly selected from the Spanish daily temperatures.

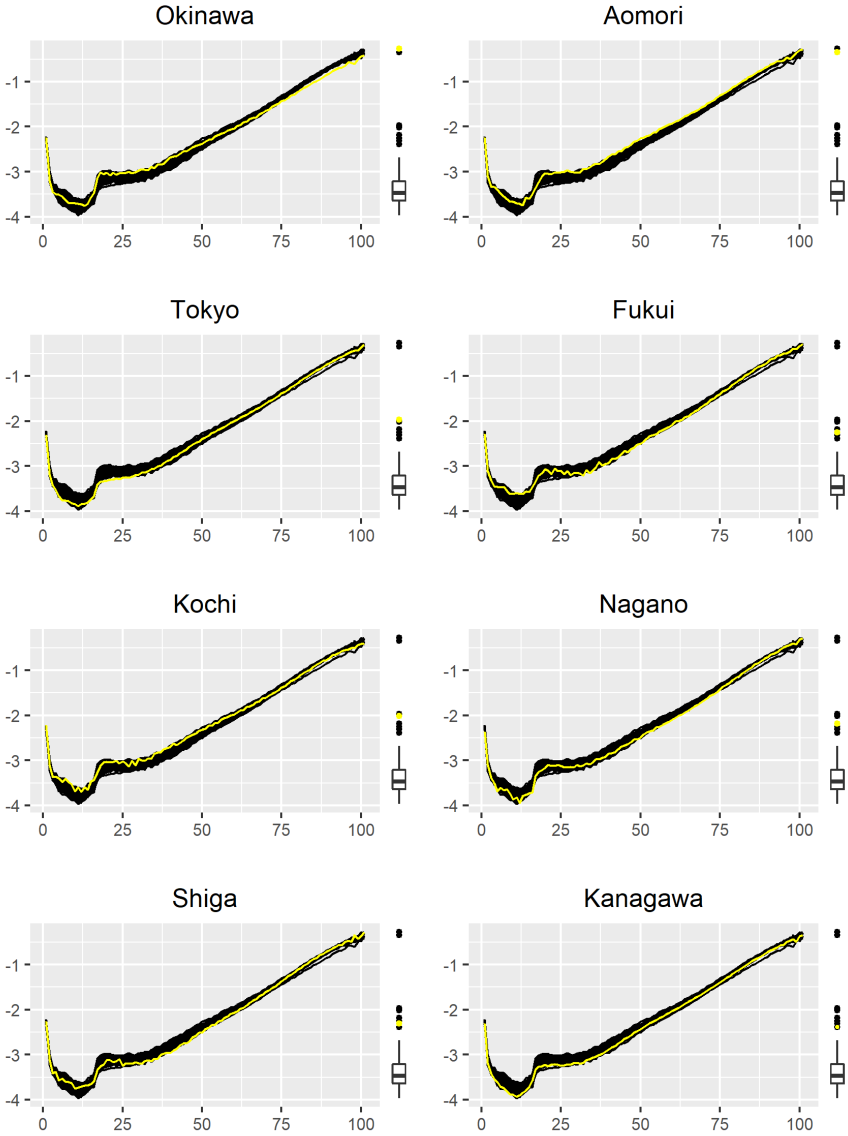

Fig. A3 Log age-specific mortality rates and localization distances boxplot for $k = 9$. Each outlier value and its corresponding curve are highlighted in yellow.

References

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

Biau B, Cérou F, Guyader A (2010) Rates of convergence of the functional k-nearest neighbor estimate. IEEE Transaction on Information Theory 56:2034–2040

Brito MR, Chávez EL, Quiroz AJ, et al. (1997) Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. Statistics & Probability Letters 35:33–42

Chen Y, Carroll C, Dai X, et al. (2020) `fdapace`: Functional Data Analysis and Empirical Dynamics. R package version 0.5.2. Development version at https://github.com/functionaldata/tPACE

Cuesta-Albertos JA, Febrero-Bande M, Oviedo de la Fuente M (2017) The DD^g-classifier in the functional setting. TEST 26(1):119–142

Dai W, Genton MG (2018) Multivariate functional data visualization and outlier detection. Journal of Computational & Graphical Statistics 27:923–934

Dai W, Genton MG (2019) Directional outlyingness for multivariate functional data. Computational Statistics & Data Analysis 131:50–65

Elías A, Jiménez R, Yukich JE (2021) localFDA: Localization Processes for Functional Data Analysis. URL https://CRAN.R-project.org/package=localFDA, R package version 1.0.0.

Elías A, Jiménez R, Shang HL (2022) On projection methods for functional time series forecasting. Journal of Multivariate Analysis 189:104,890. https://doi.org/https://doi.org/10.1016/j.jmva.2021.104890

Febrero-Bande M, Oviedo M (2012) Statistical computing in functional data analysis: The R package `fda.usc`. Journal of Statistical Software 51(4):1–28

Febrero-Bande M, Galeano P, González-Manteiga W (2019) Estimation, imputation and prediction for the functional linear model with scalar response with responses missing at random. Computational Statistics & Data Analysis 131:91–103

Ferraty F, Vieu P (2006) Nonparametric functional data analysis. Springer, New York

Gao Y, Shang HL, Yang Y (2019) High-dimensional functional time series forecasting: An application to age-specific mortality rates. Journal of Multivariate Analysis 170:232–243

Györfi L, Kohler M, Krzyzak A, et al. (2002) A Distribution-Free Theory of Nonparametric Regression. Springer, New York

Hubert M, Rousseeuw P, Segaert P (2017) Multivariate and functional classification using depth and distance. Advances in Data Analysis and Classification 11:445–466

Hyndman RJ, Booth H (2008) Stochastic population forecasts using functional data models for mortality, fertility and migration. International Journal of Forecasting 24:323–342

Hyndman RJ, Shang HL (2010) Rainbow plots, bagplots and boxplots for functional data. Journal of Computational & Graphical Statistics 19(1):29–45

Hyndman RJ, Ullah S (2007) Robust forecasting of mortality and fertility rates: a functional data approach. Computational Statistics & Data Analysis 51:4942–4956

Kara LZ, Laksaci A, Rachdi M, et al. (2017) Data-driven $k$nn estimation in nonparametric functional data analysis. Journal of Multivariate Analysis 153:176–188

Kneip A, Liebl D (2020) On the optimal reconstruction of partially observed functional data. Annals of Statistics 48:1692–1717

Kraus D (2015) Components and completion of partially observed functional data. Journal of the Royal Statistical Society: Series B-Statistical Methodology 77:777–801

Kudraszow N, Vieu P (2013) Uniform consistency of $k$nn regressors for functional variables. Statistics & Probability Letters 83:1863–1870

Li J, Cuesta-Albertos JA, Liu RY (2012) DD-classifier: Nonparametric classification procedure based on DD-plot. Journal of the American Statistical Association: Theory and Methods 107:737–753

Lian H (2011) Convergence of functional k-nearest neighbor regression estimate with functional responses. Electronic Journal of Statistics 5:31–40

Liebl D (2019) Nonparametric testing for differences in electricity prices: The case of the fukushima nuclear accident. Annals of Applied Statistics 13:1128–1146

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

Martínez F, Frías MP, Pérez MD, et al. (2017) A methodology for applying k-nearest neighbor to time series forecasting. Artificial Intelligence Review 52:2019–2037

O'Donoghue JJ (2019) Salt and inaction blamed for Aomori having the lowest life expectancy in Japan. The Japan Times Available at https://www.japantimes.co.jp/?post_type=news&p=2340547

Penrose MD (2007) Laws of large numbers in stochastic geometry with statistical applications. Bernoulli 13(4):1124–1150

Penrose MD, Yukich JE (2003) Weak laws of large numbers in geometric probability. Annals of Applied Probability 13:277–303

Ramaswamy S, Rastogi R, Shim K (2000) Efficient algorithms for mining outliers from large data sets. In: Proceedings of the ACM SIGMOD Conference on Management of Data, pp 427–438

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

Schreiber T (2010) New perspectives in stochastic geometry, Oxford University Press, Oxford, chap Limit theorems in stochastic geometry, pp 111–144

Segaert P, Hubert M, Rousseeuw P, et al. (2019) `mrfDepth`: Depth Measures in Multivariate, Regression and Functional Settings. URL https://CRAN.R-project.org/package=mrfDepth, R package version 1.0.11.

Shang HL, Hyndman RJ (2017) Grouped functional time series forecasting: An application to age-specific mortality rates. Journal of Computational & Graphical Statistics 26(2):330–343

Sun Y, Genton MG (2011) Functional boxplots. Journal of Computational & Graphical Statistics 20:316–334

Wang JL, Chiou JM, Müller HG (2016) Functional data analysis. Annual Review of Statistical Applications 3:257–295

Wu X, Kumar V, Ross Quinlan J, et al. (2008) Top 10 algorithms in data mining. Knowledge and Information Systems 14:1–37

Yao F, Müller HG, Wang JL (2005) Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association: Theory and Methods 100:577–590

Zhang S, Jank W, Shmueli G (2010) Real-time forecasting of online auctions via functional k-nearest neighbors. International Journal of Forecasting 26:666–683