Computational Statistics
https://doi.org/10.1007/s00180-022-01282-9

ORIGINAL PAPER

Depth-based reconstruction method for incomplete functional data

Antonio Elías[1] · **Raúl Jiménez**[2] · **Han Lin Shang**[3]

Received: 21 January 2022 / Accepted: 6 September 2022

Abstract

The problem of estimating missing fragments of curves from a functional sample has been widely considered in the literature. However, most reconstruction methods rely on estimating the covariance matrix or the components of its eigendecomposition, which may be difficult. In particular, the estimation accuracy might be affected by the complexity of the covariance function, the noise of the discrete observations, and the poor availability of complete discrete functional data. We introduce a non-parametric alternative based on depth measures for partially observed functional data. Our simulations point out that the benchmark methods perform better when the data come from one population, curves are smooth, and there is a large proportion of complete data. However, our approach is superior when considering more complex covariance structures, non-smooth curves, and when the proportion of complete functions is scarce. Moreover, even in the most severe case of having all the functions incomplete, our method provides good estimates; meanwhile, the competitors are unable. The methodology is illustrated with two real data sets: the Spanish daily temperatures observed in different weather stations and the age-specific mortality by prefectures in Japan. They highlight the interpretability potential of the depth-based method.

Keywords Functional data · Partially observed data · Reconstruction · Depth measures

✉ Antonio Elías
aelias@uma.es

[1] OASYS group, Department of Applied Mathematics, Universidad de Málaga, Málaga, Spain

[2] Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain

[3] Department of Actuarial Studies and Business Analytics, Macquarie University, Sydney, Australia

Published online: 26 September 2022

1 Introduction

Partially observed functional data (POFD) are becoming more recurrent, invalidating many existing methodologies of Functional Data Analysis (FDA) (Ramsay and Silverman 2005; Ferraty and Vieu 2006). Diverse case studies motivate the development of statistical tools for these data types. For example, many data sets are recorded in medical studies through periodical check-ups. Patients who miss appointments or devices that fail to register may be typical sources of censoring. These situations may present in different types of monitoring, such as ambulatory blood pressure, the health status of human immunodeficiency virus tests, growth curves, and the evolution of lung function (James et al. 2000; James and Hastie 2001; Delaigle and Hall 2013; Kraus 2015; Delaigle and Hall 2016). Sangalli et al. (2009, 2014) also consider POFD from aneurysm studies where the source of censoring comes from a prior reconstruction of the sample and posterior processing to make the data comparable across subjects. In demography, it is common that age-specific mortality rates for older ages are not completely observed due to the decreasing number of survivors (Human Mortality Database 2022) and this cohort is the focus of actuarial science studies (see, e.g., D'Amato et al. 2011). Other examples involve electricity supply functions that may not be completely observed because suppliers and buyers typically agree on prices and quantities depending on the market conditions (Kneip and Liebl 2020; Liebl and Rameseder 2019).

Prominent literature has tackled estimating missing parts of POFD, providing several benchmark methods. Among them, the methods of (Yao et al. 2005; Goldberg et al. 2014; Kraus 2015; Delaigle and Hall 2016; Kneip and Liebl 2020). This article proposes a new method and compares it with two of the above benchmark methods. We selected those that provided the best performance to minimize the mean squared prediction error (MSPE) on our simulations and considered case studies. The chosen benchmark methods are the method of Kraus (2015) and the method of Kneip and Liebl (2020). Notably, Kraus (2015) develops a procedure based on functional ridge regression with automatic parameters to predict the principal component scores and the unobserved part of a function when only a fragment of the curve is available. In Kraus and Stefanucci (2020), the authors prove that this ridge reconstruction method is asymptotically optimal. On the other hand, Kneip and Liebl (2020) approaches the problem by introducing a new optimal operator based on a local linear kernel to produce smooth results and avoid artificial jumps between observed and reconstructed parts. The best-performing reconstruction methods that outperform Yao et al. (2005) and Kraus (2015) involve an alignment step to link the predicted fragments with the partially observed curve. In this article, we focus on a comparison between methods without any post-processing step.

Our approach combines two novel functional tools: (1) the concept of functional envelope and projection methods proposed in Elías et al. (2022b); and (2) a functional depth for POFD (Elías et al. 2022a). The literature on depth-based reconstruction methods for functional data is scarce. Remarkably, Mozharovskyi et al. (2020) proposed an iterative imputation method for multivariate data by data depth motivated by iterative regression imputation methods. The authors propose to replace

the missing values on some of the dimensions with the corresponding value that maximizes the depth on the remaining completely observed dimensions. Differently, we select a subset of sample trajectories representative of the curve to reconstruct in terms of shape and magnitude, called envelope (Elías et al. 2022b), that maximizes the functional depth of the curve to reconstruct. Then, we "project" the envelope to the missing regions and use all the projected curves to provide a reconstruction. We do not restrict our search to a completely observed part of the domain nor to completely observed functions, thanks to a suitable functional depth for partially observed data (Elías et al. 2022a) and a measure of similarity between POFD. In our context, every single datum might contain missing parts, and we do not require to have complete observability at any region of the curves' domain. This fact also makes our problem different from the imputation problem of Liebl and Rameseder (2019) where the missing parts are not systematic.

The new method is designed for: (1) Scenarios where the estimation of the covariance function is complex or impossible: The existing reconstruction methods depend on a proper covariance estimation. Its estimators might be very sensitive and become unreliable for many analyses, particularly for principal component analysis (Hubert et al. 2005). In addition, data coming from multiple populations and low number of complete functions in the sample also hampers the estimation procedures and makes it impossible if all the sample functions are partially observed (Kneip and Liebl 2020). (2) Reconstructing non-smooth functions: Some methods are designed to deal with smooth functions; consequently, their results are smooth and aligned functions (Kneip and Liebl 2020). Our goal is to provide a method that produces reconstructed functions while remaining as accurate as possible in roughness and variability. (3) Adding interpretability: It might be useful to get a precise estimation and insights into the final reconstruction drivers. We consider simulated and empirical data for illustrating the issues listed above. On the one hand, we consider yearly curves of Spanish daily temperatures. This data set is gathered by the Spanish Agency of Meteorology (AEMET) at different weather stations spread along with the Spanish territory. On the other hand, we consider age-specific yearly mortality rates recorded at each Japanese prefecture (political territory division).

The structure of this paper is as follows: Sect. 2 introduces notation and the method. Section 3 shows various simulated results based on Gaussian Processes under different simulated regimes of partial observability. In addition, we illustrate the method's performance with the Spanish daily temperatures data set and the Japanese age-specific mortality rates by prefecture. In Sect. 4, we make some conclusions, along with some ideas on how the methodology can be further extended.

2 Depth-based reconstruction method

2.1 Definition and notation

Let $X = \{X(t) : t \in [a, b]\}$ be a stochastic process of continuous trajectories and $(X_1, \ldots, X_n)$ independent copies of X. To simplify the notation, we assume without loss of generality $[a, b] = [0, 1]$. We consider the case $X_1, \ldots, X_n$ are partially

observed. Following Delaigle and Hall (2013), we model the partially observed setting by considering a random mechanism Q that generates compact subsets of [0, 1] where the functional data are observed. Specifically, let O be a random compact set generated by Q, and let $(O_1, \ldots, O_n)$ be independent copies of O. Therefore, for $1 \leq i \leq n$, the functional datum X_i is only observed on O_i. Let $(X_i, O_i) = \{X_i(u) : u \in O_i\}$ and $M_i = [0, 1] \setminus O_i$. Then, the observed and missing parts of X_i are (X_i, O_i) and (X_i, M_i). As it is standard in the literature of POFD, we assume that $(X_1, O_1), \ldots, (X_n, O_n)$ are i.i.d. realizations from $P \times Q$. This is, $\{X_1, \ldots, X_n\}$ and $\{O_1, \ldots, O_n\}$ are independent samples. This assumption has been termed Missing-Completely-at-Random and, notedly, only Liebl and Rameseder (2019) has considered a specific violation of this assumption. As illustration, the top panel of Fig. 1 presents two examples of incomplete functional data under two different missing scenarios (left and right panels). The observed fragments of two incomplete curves (X_i, O_i) are in red.

The core idea of our method is based on the depth-based method for dynamic updating on functional time series Elías et al. (2022b). In this context, by using the notation introduced here, the functional sample, $(X_1, \ldots, X_n)$, was ordered in time, n being the most recent time period. X_n was only observed on $O_n = [0, q]$, with $0 \ll q < 1$, and the rest of sample curves $\{X_j : j < n\}$ were fully observed on [0, 1]. This is, for all $j < n$, $O_j = [0, 1]$. We may summarize the depth-based approach for dynamic updating as follows: if (X_n, O_n) is depth in $\{(X_j, O_n) : j \in \mathcal{J}_n\}$, for some set of curves $\mathcal{J}_n \subset \{1, \ldots, n-1\}$, and the band delimited by the curve segments $\{(X_j, O_n) : j \in \mathcal{J}_n\}$ captures both the shape and magnitude of (X_n, O_n), we may estimate (X_n, M_n) from $\{(X_j, M_n) : j \in \mathcal{J}_n\}$. In particular, point estimators of (X_n, M_n) were obtained by computing weighted averages on the curve segments of $\{(X_j, M_n) : j \in \mathcal{J}_n\}$. In our jargon, $\{(X_j, O_n) : j \in \mathcal{J}_n\}$ is called the *envelope* of (X_n, O_n).

In contrast to the dynamic updating framework of Elías et al. (2022b), this article deals with sample curves that might not be temporarily ordered in the partially observed scenario. What is more important, every single curve may be partially observed. So, for enveloping the observed part of a curve, say us (X_i, O_i), we could only have curve segments, namely $\{(X_j, O_i \cap O_j) : j \neq i\}$, with $|O_i \cap O_j| \neq \emptyset$. Moreover, without loss of generality, we assume that $O_i \cap O_j$ has a non-zero Lebesgue measure. Similarly, we may only have curve segments, specifically $\{(X_j, M_i \cap O_j) : j \neq i\}$, for estimating the missing part of X_i, that is (X_i, M_i). Applying the depth-based approach to these cases is a challenging and open problem we address in this article. The method can be explained in two steps: one concerned about how to obtain an *envelope* of each (X_i, O_i), described in Sect. 2.2 and one on how to estimate or reconstruct (X_i, M_i) from the observed parts of the curves used for enveloping (X_i, O_i), described in Sect. 2.3.

2.2 A depth-based algorithm for focal-curve enveloping

The concept of depth arises for ordering multivariate data from *the center to outward* (Liu 1990; Liu et al. 1999; Rousseeuw et al. 1999; Zuo and Serfling 2000;

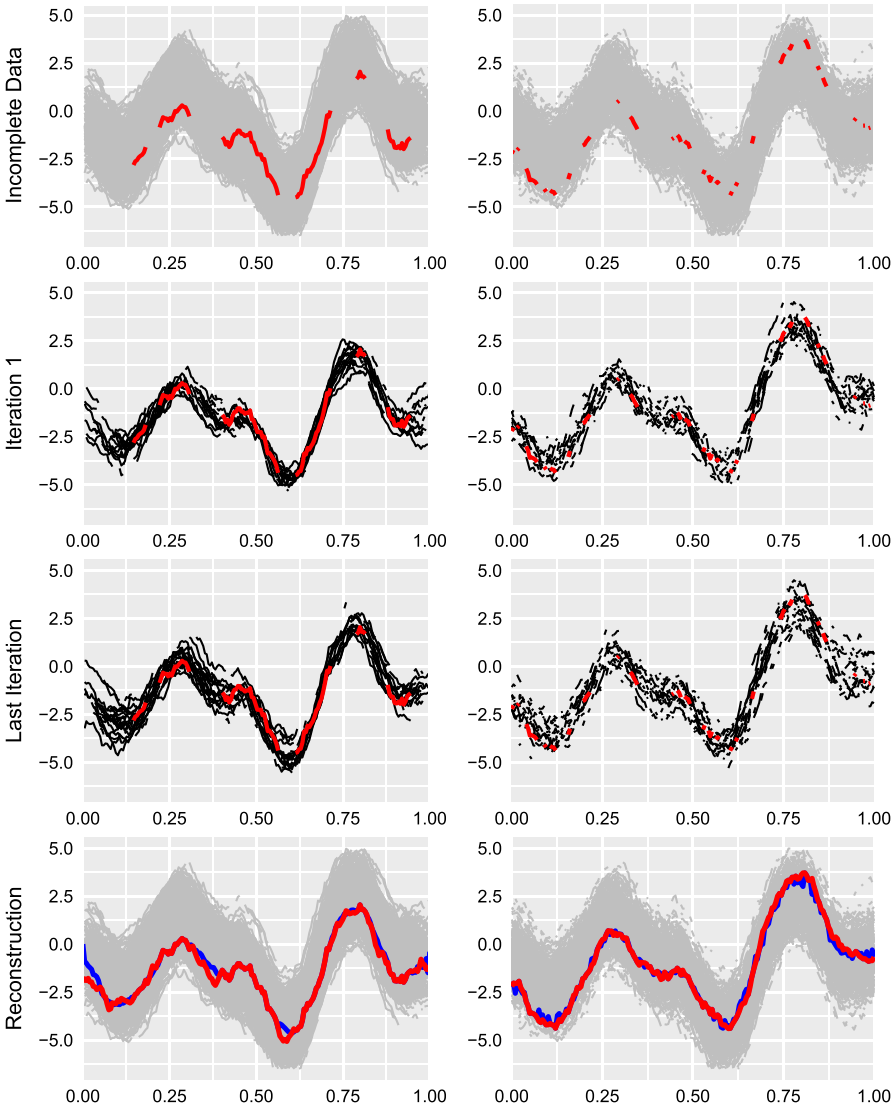

Fig. 1 In the top panels, an illustration of how Algorithm 1 works. The sample curves correspond to 1000 i.i.d. trajectories of a Gaussian process. We considered partially observed curves for the left panels by removing six random intervals from [0, 1]. For the right panels, we considered missing data uniformly. On average, only 50% of each curve was observed for both runs. The partially observed function that we reconstructed is colored in red and plotted entirely in the bottom panels jointly with its estimation (in blue)

Serfling 2006; Li et al. 2012). Let $\mathcal{F}$ be the collection of all probability distribution functions on $\mathbb{R}$, $F \in \mathcal{F}$ and $x \in \mathbb{R}$. In the univariate context, a depth measure is a function $D : \mathbb{R} \times \mathcal{F} \to [0, 1]$ such that, for any fixed F, $D(x, F)$ reaches their maximum value at the median of F, this is at x such that $F(x) = 1/2$, and decreases

to the extent that x is farthest from the median. Examples of such univariate depth measures are

$$D(x, F) = 1 - \left|\frac{1}{2} - F(x)\right| \tag{1}$$

and

$$D(x, F) = 2\{F(x)\big[1 - F(x)\big]\}. \tag{2}$$

Denote by P to the generating law of the process X and by P_t to the marginal distribution of $X(t)$, this is $P_t(x) = \mathbb{P}[X(t) \leq x]$. Given a univariate depth measure D, the Integrated Functional Depth of X with respect to P is defined as

$$\text{IFD}\,(X, P) = \int_0^1 D\big(X(t), P_t\big) w(t) dt, \tag{3}$$

where w is a weight function that integrates to one (Claeskens et al. 2014; Nagy et al. 2016). It is worth mentioning that when $w(t) = 1, \forall t$, and D is defined by (1), the integrated functional depth corresponds to the seminal Fraiman and Muniz 's (2001) functional depth. When D is defined by (2), the corresponding IFD is the celebrated Modified Band Depth with bands formed by two curves (López-Pintado and Romo 2009).

Define $\mathcal{J}(t) = \{1 \leq j \leq n : t \in O_j\}$. Suppose $\mathcal{J}(t) \neq \emptyset$ and let $q(t)$ be the cardinality of $\mathcal{J}(t)$. Denote by $F_{\mathcal{J}(t)}$ to the empirical distribution function of the univariate sample $\{X_j(t) : j \in \mathcal{J}(t)\}$. This is the probability distribution that assigns constant mass equals to $1/q(t)$ to each available observation at time t. Then, for any pair (X_i, O_i) and a given univariate depth D, we consider the Partially Observed Integrated Functional Depth (Elías et al. 2022a) restricted to O_i defined by

$$\text{POIFD}\left((X_i, O_i), P \times Q\right) = \frac{\int_{O_i} D\big(X(t), F_{\mathcal{J}(t)}\big) q(t) dt}{\int_{O_i} q(t) dt}. \tag{4}$$

This definition of depth considers that the sample of curves is incomplete and weights the parts of the domain proportionally to the number of observed curves. The larger $\text{POIFD}\,((X_i, O_i), P \times Q)$, the deeper will be (X_i, O_i) in the partially observed sample.

In line with the approach for dynamic updating introduced by Elías et al. (2022b), we search for an envelope $\mathcal{J}$ that is a subset of incomplete curves, as big as possible, that captures the shape and magnitude of (X_i, O_i) such that $i \notin \mathcal{J}$ and with the following desirable properties:

P1) (X_i, O_i) is deep in $\{(X_j, O_j) : j \in \mathcal{J} \cup \{i\}\}$, the deepest if possible. For measuring depth here we use the Partially Observed Integrated Functional Depth restricted to O_i defined in (4).

P2) (X_i, O_i) is enveloped by $\{(X_j, O_j) : j \in \mathcal{J}\}$ as much as possible. Here, we say (X_i, O_i) is *more enveloped* by $\{(X_j, O_j) : j \in \mathcal{J}\}$ than by $\{(X_j, O_j) : j \in \mathcal{J}'\}$ if and only if

$$\lambda\left(\left\{t \in O_i : \min_{j\in\mathcal{J}(t)} X_j(t) \le X_i(t) \le \max_{j\in\mathcal{J}(t)} X_j(t)\right\}\right) > \lambda\left(\left\{t \in O_i : \min_{j\in\mathcal{J}'(t)} X_j(t) \le X_i(t) \le \max_{j\in\mathcal{J}'(t)} X_j(t)\right\}\right),$$

with λ being the Lebesgue measure on $\mathbb{R}$.

P3) $\{(X_j, O_j) : j \in \mathcal{J}\}$ contains near curves to (X_i, O_i), as many as possible. For measuring nearness, we use *mean* L_2 *distance* between couples of POFD that overlap i.e. $\lambda(O_i \cap O_j) > 0$. This is,

$$\|(X_i, O_i) - (X_j, O_j)\| = \frac{\sqrt{\int_{O_i\cap O_j} |X_i(t) - X_j(t)|^2 dt}}{\lambda(O_i \cap O_j)}. \tag{5}$$

Algorithm 1 provides a set of curves with the three features above that we call the *i-curve envelope* and denote by $\mathcal{J}_i$ hereafter. The algorithm is a variation of Algorithm 1 of Elías et al. (2022b), adapted to POFD. It iteratively selects as many sample curves as possible, from the nearest to the farthest to (X_i, O_i), for enveloping (X_i, O_i) (algorithm lines from 3 to 10) and increasing its depth at each iteration (algorithm lines from 11 to 13). The second row of Fig. 1 presents the first iteration where the black curves are not only the closest ones to their corresponding red curve (P3) but also they surround and cover the curve to reconstruct (X_i, O_i) on O_i (P2). The third row is the last iteration of the algorithm that contributes with an additional set of curves that makes (X_i, O_i) deeper (P1).

Algorithm 1: Envelope for partially observed functional data

Input: $i, \{(X_j, O_j) : 1 \le j \le n\}$.
Output: Envelope $\mathcal{J}$

1 **Initialize** $f = (X_i, O_i), \mathcal{Y} = \{(X_j, O_i \cap O_j) : j \ne i\}, \mathcal{J} = \emptyset$ and $D(f|\mathcal{J}, i) = 0$ **while** *size of* $\mathcal{Y} \ge 2$ **do**
2 Let y' be the nearest curve to f from $\mathcal{Y}$ and $\mathcal{N} = \{y'\}$
3 **for** $y \in \mathcal{Y} \setminus \{y'\}$, *from the nearest curve to the farthest from* f, **do**
4 $j_y = \{j : (X_j, O_i \cap O_j) = y\}$
5 $\mathcal{J}^+ = \mathcal{J} \cup \{j : (X_j, O_i \cap O_j) \in \mathcal{N}\}$
6 **if** f *is more enveloped by* $\mathcal{N} \cup \{y\}$ *than by* $\mathcal{N}$ *or* $O_{j_y} \setminus \left(\cup_{j\in\mathcal{J}^+} O_j\right) \ne \emptyset$ **then**
7 $\mathcal{N} = \mathcal{N} \cup \{y\}$
8 **end**
9 **end**
10 **if** $D(f|\mathcal{J} \cup \mathcal{N}, i) \ge D(f|\mathcal{J}, i)$ **then**
11 $\mathcal{J} = \mathcal{J} \cup \{j : (X_j, O_i \cap O_j) \in \mathcal{N}\}$
12 **end**
13 $\mathcal{Y} = \mathcal{Y} \setminus \mathcal{N}$
14 **end**

Figure 1 illustrates how Algorithm 1 works. We consider the first, second, and final iteration of two runs from the algorithm based on 1000 i.i.d. trajectories of a Gaussian process. We considered partially observed curves for the run shown in the left panels by removing six random intervals of the observation domain for every single functional datum. We considered missing data uniformly on the observation domain for the run shown in the right panels. On average, only 50% of each curve was observed for both runs. The partially observed function we intend to reconstruct is colored in red and plotted entirely in the bottom panels jointly with its estimation that we describe below.

2.3 Reconstruction of missing parts

For estimating the unobserved part of X_i, this is (X_i, M_i), we use a weighted functional mean from data of the curve envelope $\mathcal{J}_i$. Only here, these functional data may be partially observed. Specifically, these data are $\{(X_j, M_i \cap O_j) : j \in \mathcal{J}_i\}$. Consider $\mathcal{J}_i(t) = \{j \in \mathcal{J}_i : t \in O_j\}$, assume $\mathcal{J}_i(t) \neq \emptyset$ for all $t \in O_i$, and let $\delta = \min_{j \in \mathcal{J}_i} \|(X_i, O_i) - (X_j, O_j)\|$. Then, we estimate X_i on M_i by

$$\widehat{X}_i^{\theta}(t) = \frac{\sum_{j \in \mathcal{J}_i(t)} w_j X_j(t)}{\sum_{j \in \mathcal{J}_i(t)} w_j}, \quad \text{with} \quad w_j = \exp\left(\frac{-\theta \|(X_i, O_i) - (X_j, O_j)\|}{\delta}\right). \tag{6}$$

This estimator is a version of the envelope projection with exponential weights (see Elías et al. 2022b, Equation (2)), adapted to the partially observed data context. Notice that the exponential weights increase the influence of the closest curves in the estimation and give little importance to the farthest trajectories of the envelope. The parameter θ is automatically chosen by minimizing the mean squared error (MSE) on (X_i, O_i), and it tunes the importance of each curve of the envelope in the reconstruction. In practice, if $\widehat{O}_i$ is the observational set where $\widehat{X}_i$ can be computable, this is $\cup_{j \in \mathcal{J}_i}(O_i \cap O_j)$, then

$$\theta = \arg\min_{\nu} \sum_{i=1}^{n} \|(X_i, O_i) - (\widehat{X}_i^{\nu}, \widehat{O}_i)\|^2.$$

As an illustration, the bottom panels of Fig. 1 show reconstructions of missing parts of the two simulated cases discussed above.

3 Results

We compare results obtained using the depth-based method with those obtained from studies by Kraus (2015) and Kneip and Liebl (2020). Kraus (2015) propose a regularized regression model to predict the principal component scores (Reg. Regression), whereas Kneip and Liebl (2020) introduces a new class of reconstruction operators that are optimal (Opt. Operator). The two methods were implemented

by using the R-codes available at https://is.muni.cz/www/david.kraus/web_files/papers/partial_fda_code.zip and https://github.com/lidom/ReconstPoFD. The depth for partially observed curves and the data generation settings are implemented using the R-package `fdaPOIFD` of Elías et al. (2021).

Section 3.1 introduces the simulation setting and the data generation process for POFD and shows results with synthetic data. Section 3.2 uses the same simulation settings but applied to AEMET temperature data. Additionally, it illustrates the reconstruction of some yearly temperature curves that are partially observed in reality. Finally, Sect. 3.3 presents another real case study where Japanese age-specific mortality functions are reconstructed.

3.1 Simulation study

Let us denote by $c\%$ the percentage of sample curves that are partially observed. Benchmark methods perform better as the parameter c is larger. This finding is because these reconstruction methods strongly depend on the information of the completely observed curves to estimate the covariance or the components of its eigendecomposition. However, the depth-based method can handle the case $c = 0$, i.e., there are no complete functions in the sample. Therefore, results for this case are reported without comparison.

We considered two Missing-Completely-at-Random procedures for generating partially observed data for our simulation study. These procedures have previously been used in the literature (Elías et al. 2022a) and are in line with the partial observability of the real case studies. They are:

Random Intervals, with which $c\%$ of the sample curves is observed on a number m of random disjointed intervals of $[0, 1]$.

Random points, with which $c\%$ of the functions is observed on a very sparse random grid.

First, we apply these observability patterns to simulated trajectories. Concretely, we consider a Gaussian process $X(t) = \mu(t) + \epsilon(t)$ where $\epsilon(t)$ is a centered Gaussian process with covariance kernel $\rho_\epsilon(s,t) = \alpha e^{-\beta|s-t|}$ for $s, t \in [0, 1]$. The functional mean $\mu(t)$ is a periodic function randomly generated by a centered Gaussian process with covariance $\rho_\mu(s,t) = \sigma e^{-(2\sin(\pi|s-t|)^2/l^2)}$. Thus, each sample will present different functional means. The set of parameters used for our study were $\beta = 2$, $\alpha = 1$, $\sigma = 3$ and $l = 0.5$. Examples of the generated trajectories by this model are those shown in Fig. 1.

We considered small and large sample sizes for the study by making $n = 200$ and 1000. Also, we considered different percentages of observed curves that were partially observed. Specifically, we tested with $c = 0, 25, 50$ and 75. In addition, we considered different percentages of time on which the incomplete curves of a

sample were observed. Henceforth, we term this percentage by $p\%$. We considered $p = 25, 50$ and 75 for small samples but only $p = 25$ and 50 for large samples. This is due to the computational cost of the benchmark methods when $n = 1000$ and $p = 75$. Note that this parameter setting implies the highest computational cost for estimating covariance functions. Finally, we replicate 100 samples of each data set to estimate median values of MSPE.

Table 1 presents results for $n = 200$. It shows MSPE from the Gaussian data and points out the superiority of the Reg. Regression method (Kraus 2015) when covariance function is simple to estimate, as is the case of the exponential decay covariance function involved in these data (see left panel of Fig. 2). Even in this case, we remark that the depth-based method is slightly better than the Opt. Operator method (Kneip and Liebl 2020). When all the functions of the sample are partially observed ($c = 0\%$), only the depth-based method can provide a reconstruction, and, surprisingly, the MSPE remains reasonably similar to those cases with a proportion of complete functions significantly large ($c = 25, 50, 75\%$). With regards to estimation uncertainty, the depth-based method is superior to Opt. Operator and comparable with Reg. Regression, observed from the top panel of Fig. 3. Similar results are obtained with other sample sizes and the Random Interval setting (see the detailed results in Appendix 1.1).

Appendix 1.2 includes results with functional data generated from truncated Karhunen-Loève expansions following the simulations in Liebl and Rameseder (2019) for sample sizes of 100 and 500. First, we consider one population data and, in this setting, Opt. performs better than the Gaussian processes making Opt. Operator method and Reg. Regression comparable. This finding is in line with the results reported in Liebl and Rameseder (2019). Additionally, motivated by our empirical case studies, we consider a setting with smooth functions generated with truncated Karhunen-Loève expansions for generating multiple populations. In this context, our depth-based proposal starts being competitive, achieving better results than Opt. Operator and sometimes even better than Reg. Regression. The detailed explanation of the data generation and the results are reported in Sect. A.3.

3.2 Case study: reconstructing AEMET temperatures

Spanish Agency of Meteorology (AEMET) provides meteorological variables recorded from different stations in the whole Spanish territory (see http://www.aemet.es/es/portada). This analysis focus on maximum daily temperatures of 73 stations located in the capital of provinces. Following the literature of FDA, we consider this data as a functional data set where each function is the temperatures of each complete year (see also Febrero-Bande and Oviedo de la Fuente 2012; García-Portugués et al. 2014). Some of the curves are partially observed in the historical data, and our goal is to reconstruct the data set.

Temporal data availability depends from one station to the other. For example, Madrid-Retiro station is the oldest, being monitored from 1893, and Ceuta from 2003. We consider a set of 2786 entirely observed curves of different years and weather stations. This large sample of complete functions allows reproducing the

Table 1 Median values of MSPE over 100 pseudo-random replicates

Method	$c = 75$			$c = 50$			$c = 25$			$c = 0$		
	$p = 25$	50	75	25	50	75	25	50	75	25	50	75
Depth-based	0.153	0.138	0.133	0.169	0.144	0.137	0.197	0.154	0.14	**0.259**	**0.168**	**0.143**
Opt. operator	0.159	0.16	0.247	0.179	0.191	0.246	0.187	0.185	0.256	–	–	–
Reg. regression	**0.059**	**0.054**	**0.049**	**0.075**	**0.07**	**0.06**	**0.124**	**0.111**	**0.086**	–	–	–

Each replicate is composed of 200 curves. A dash (-) represents that the method cannot produce any reconstruction. The partially observed samples are obtained by observing $p\%$ of the total discrete realization points (Random Points). The smallest error is bolded for each combination of c and p

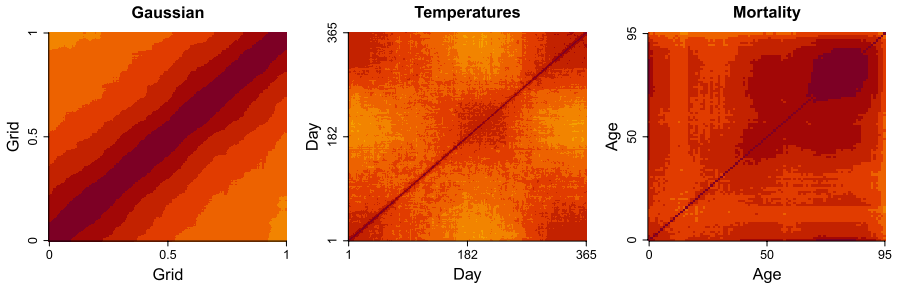

Fig. 2 Covariance estimations based on the available functions are completely observed. Left panel, Gaussian processes with an exponential decay covariance. Central panel: Spanish daily temperatures with lower covariance values in Spring and Autumn periods and higher covariance in Summer and Winter. Right panel: Japanese age-specific mortality rates with higher correlations at the oldest ages

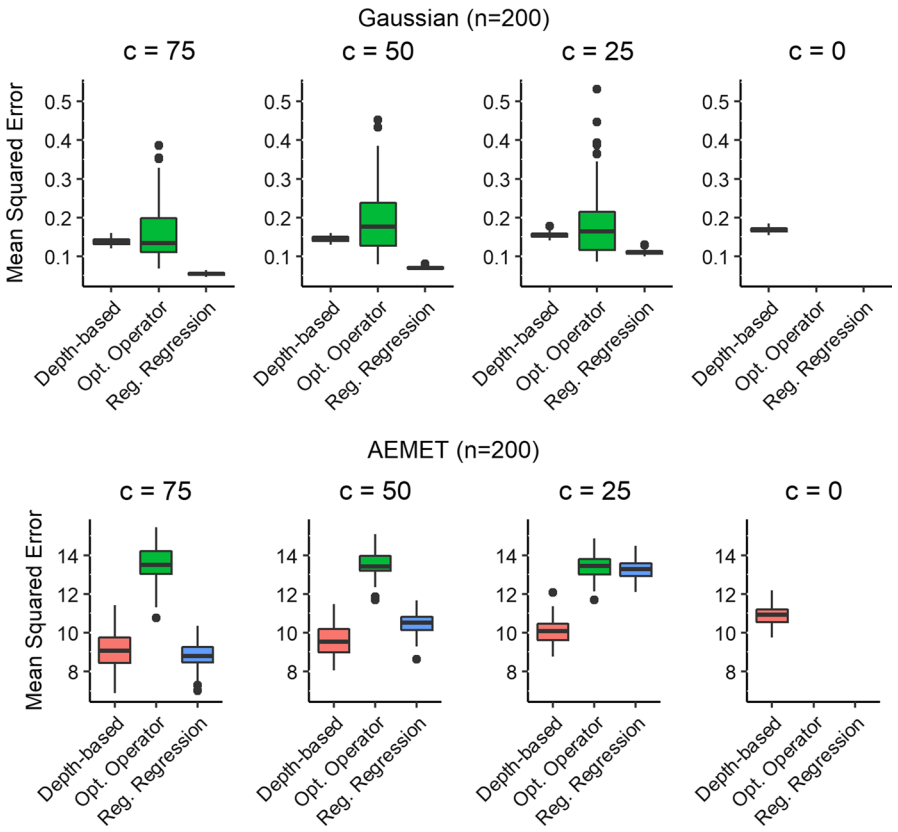

Fig. 3 Boxplots of the MSPE for 100 replicates, $p = 50$ and $n = 200$. Top (Gaussian data) and bottom (AEMET data) panels were obtained with the same data as Tables 1 and 2, respectively

simulation in Sect. 3.1 by randomly generating random functional samples. To do that, we randomly generate 100 samples of curves without replacement. The results for sample sizes of $n = 200$ and $n = 1000$ are given in Table 2. Unlike Gaussian

Table 2 Median values of MSPE over 100 pseudo-random replicates

Method	$c = 75$		$c = 50$		$c = 25$		$c = 0$	
	$p = 25$	50	25	50	25	50	25	50
Depth-based	**5.393**	**4.920**	**6.384**	**5.274**	**7.874**	**6.321**	**10.252**	**7.606**
	(9.639)	(9.099)	(10.879)	(**9.585**)	(**12.555**)	(**10.079**)	(**14.691**)	(**10.900**)
Opt. operator	13.081	13.403	13.201	13.442	13.259	13.229	–	–
	(13.184)	(13.530)	(13.296)	(13.493)	(13.412)	(13.436)	–	–
Reg. regression	6.834	5.41	7.349	5.944	8.617	8.217	–	–
	(**9.508**)	(**8.826**)	(10.842)	(10.472)	(13.577)	(13.251)	–	–

Each replicate comprises 1000 and 200 curves (results between parenthesis). A dash (-) represents that the method cannot produce any reconstruction. The partially observed samples are obtained by observing $p\%$ of the total discrete realization points (Random Points). The smallest error is bolded for each combination of c and p

data, the depth-based method is superior to the competitors in the simulation with AEMET data, showing small MSPE and comparable variances (see bottom panel of Fig. 3 for an insight into the uncertainty of the estimation for each method). Only for high proportions of complete functions $c = 75$ and small sample size $n = 200$, Kraus 's (2015) method was superior. The depth-based method is superior for smaller values of c or larger sample sizes $n = 1000$. These results can be explained by the complex structure of AEMET data and the availability of multiple stations with different weather conditions, as shown in the center panel of Fig. 2.

Table 3 shows the same simulation setup with AEMET data but under the Random Interval setting and small sample size. In this setting, we generate partially observed data for a different number of observed intervals (m), percentages of completely observed curves (p), and the mean observability percentage of each partially observed curve (c). The result shows that when $p = 25$, the regression method performs better than the competitors. This finding is because our implementation of the partially observed mechanism produces a sample of POFD with more proportion of observed curves in the center of the domain than in the extreme. Then, for small p the reconstruction of POFD observed in the middle of the domain worsens the results of the depth-based method. However, when p increases, our implementation of the random censoring mechanism produces a sample of POFD that uniformly covers the complete domain.

In Fig. 4, we plot the reconstructions obtained by the three methods under consideration from one random sample. This was obtained by randomly taking 1000 curves from the total observed curves of the AEMET data. Then, we generated partially observed data by applying the Missing-Completely-at-Random procedure based on random intervals described above, with $m = 4$, $p = 50$, and $c = 50$. Finally, we randomly selected one function to reconstruct, namely, VALLADOLID/VILLANUBLA-1956, where VALLADOLID/VILLANUBLA refers to the location of the station and 1956 is the observation year. VALLADOLID/VILLANUBLA-1956 is plotted in red according to a general view of its shape. In contrast, the reconstructions are only plotted on the four intervals where the curve was observed in our

Table 3 Median values of the MSPE after 100 pseudo-random replicates

Intervals	Method	$c = 75$			$c = 50$			$c = 25$		
		$p = 25$	50	75	25	50	75	25	50	75
$m = 1$	Depth-based	**13.359**	**10.037**	**8.999**	15.597	**11.224**	**9.690**	18.122	**12.903**	10.884
	Opt. operator	16.041	12.862	10.998	16.105	13.119	10.979	**16.567**	13.366	11.168
	Reg. regression	14.420	11.033	9.472	**15.466**	12.359	10.290	16.821	14.194	**10.877**
$m = 2$	Depth-based	13.546	**10.363**	**8.975**	16.935	**11.076**	**9.621**	19.415	**12.347**	**10.385**
	Opt. operator	16.657	13.689	11.200	16.856	13.591	11.373	16.960	13.662	11.549
	Reg. regression	**13.369**	11.394	9.542	**14.439**	12.393	10.597	**16.074**	14.228	12.422
$m = 4$	Depth-based	13.841	**10.316**	**9.199**	16.403	**10.829**	**9.579**	18.682	**11.800**	**10.097**
	Opt. operator	16.164	13.665	12.126	16.088	13.689	11.953	16.105	13.563	11.962
	Reg. regression	**12.909**	11.320	10.022	**13.676**	12.354	11.063	**15.427**	14.301	12.736

This table summarizes the exercise considering random samples from the fully observed AEMET data set. Each replicate is composed of 200 functional observations. The partially observed samples are obtained by restricting each function to m intervals of total length $p\%$ of the domain (Random Intervals). Thesmallest error is bolded for each combination of c, m, and p

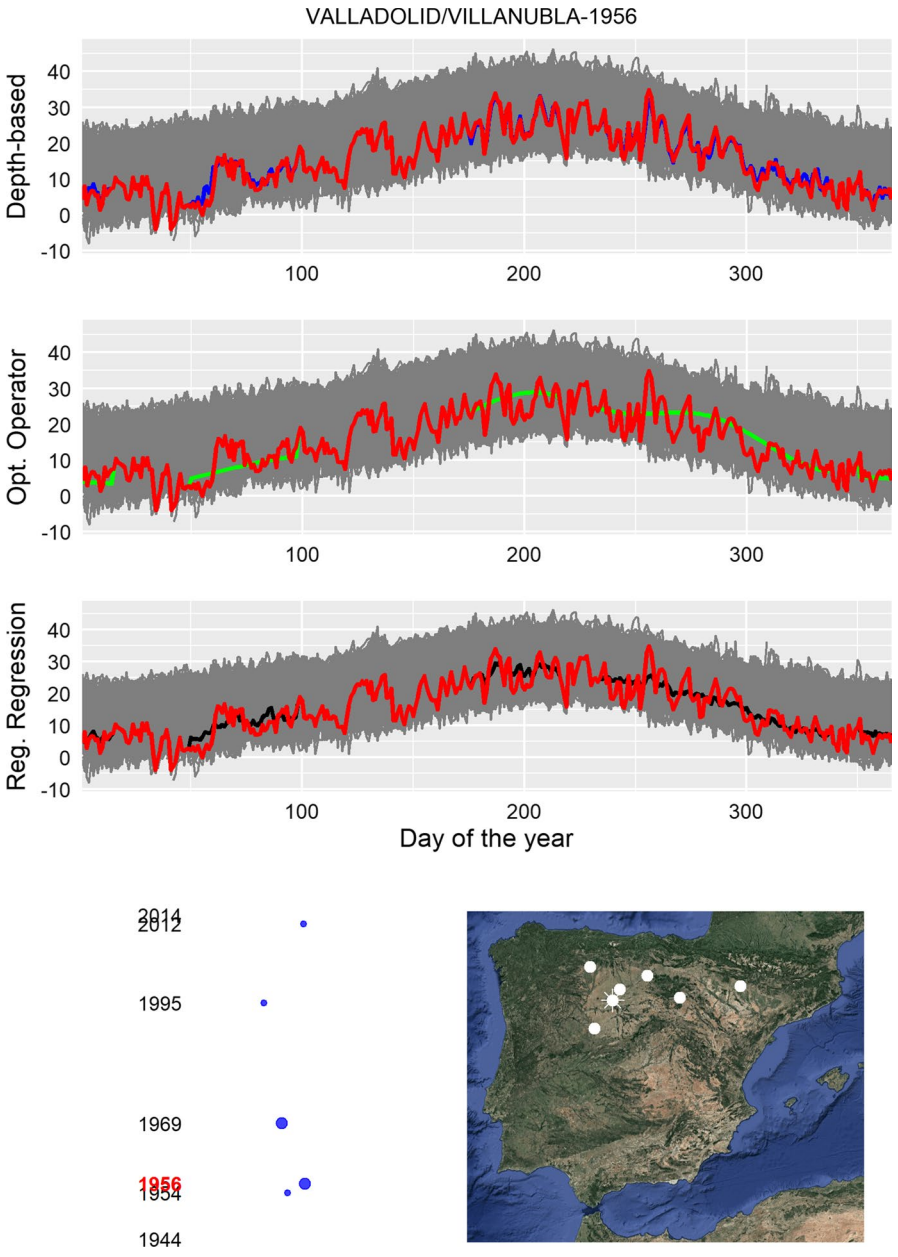

Fig. 4 Simulated exercise reconstruction of "VALLADOLID/VILLANUBLA-1956". Top three panels present the reconstruction by the depth-based method, Kraus (2015) and Kneip and Liebl (2020). Bottom panel, the spatial (Spanish map), and temporal (bubble plot) descriptive analyses are shown by the depth-based methodology and the envelope

simulation. The top panels of the figure show output produced by the depth-based method (Depth-based in blue), Kneip and Liebl (2020) (Opt. Operator in green), by Kraus (2015) (Reg. Regression in black). The depth-based method is superior to the benchmark methods. The bottom panel of the figure shows some descriptive statistics related to the depth-based method. We show the years used for reconstructing (the years of the curves in the envelope). The frequency of each year (number of curves into the envelope with the same year) is represented by a proportional blue bubble. Similarly, we show the locations of the curves on the right side of the envelope. From our understanding, all of them have similar geographical features.

Finally, Fig. 5 illustrates the actual case of "BURGOS/VILLAFRÍA-1943" a station that probably started operating in the middle of the year 1943. Consequently, only the year's second half is recorded (red curve at the top panel). We apply the

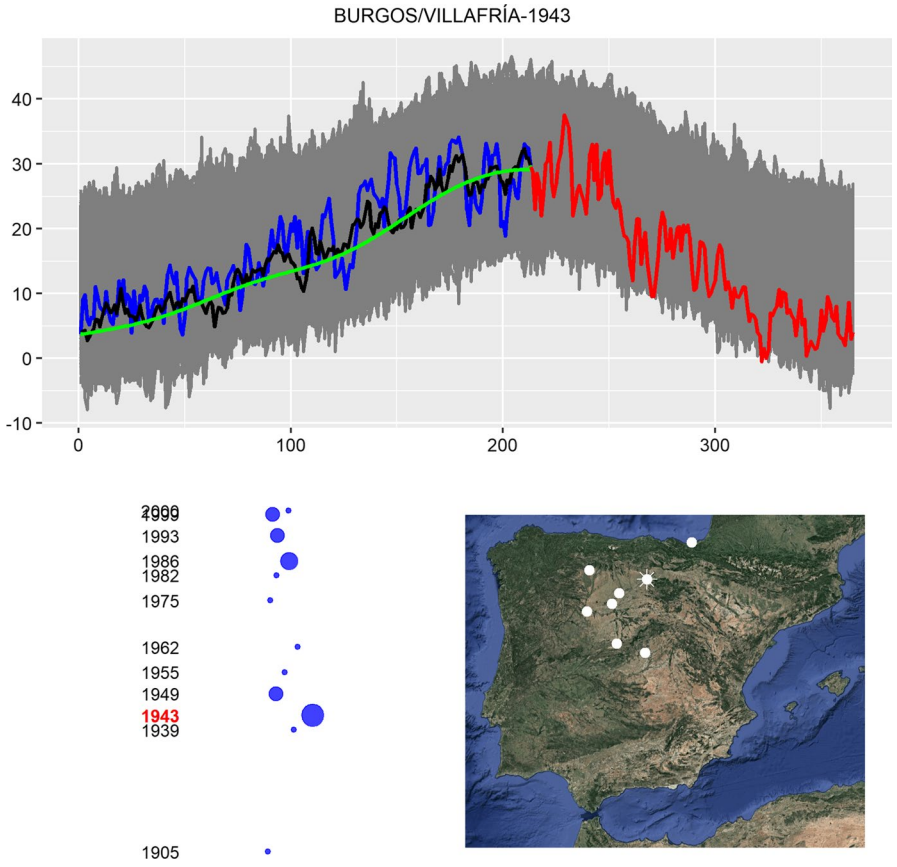

Fig. 5 Holdout partially observed function, "BURGOS/VILLAFRÍA-1943". Top panel: reconstructions by Kraus (2015) (in black) Kneip and Liebl (2020) (in green) and the depth-based method (in blue). The bottom panel shows the descriptive analysis of the most relevant curves in reconstructing "BURGOS/VILLAFRÍA-1943". The left part is time analysis (bubble plot of the involved years); the right is spatial analysis (map with the most relevant and involved stations)

three reconstructing methods to complete the first half of the curve (Reg. Regression (Kraus 2015) in black, Opt. Operator (Kneip and Liebl 2020) in green, and the depth-based method in blue). The depth-based methodology supports the reconstruction with the additional information provided by the most important curves of the envelope. In this case, the envelope of "BURGOS/VILLAFRÍA-1943" contains distant-past functions from 1905 from the MADRID-RETIRO station and also more recent functions from the 90s from the same station. Additionally, the largest proportion of functions belongs to 1943 (biggest blue bubble), the same year of the curve to reconstruct.

3.3 Case study: reconstructing Japanese mortality

The Human Mortality Data Set (https://www.mortality.org) provides detailed mortality and population data of 41, mainly in developed countries. Some countries also offer micro-information by subdividing territory, providing challenging spatial and temporal information. In particular, the Japanese mortality data set is available at http://www.ipss.go.jp/p-toukei/JMD/index-en.asp for its 47 prefectures for males, females, and the total population.

A common FDA approach to analyzing mortality data is to consider that each function is the yearly mortality for each age cohort (see, e.g. Shang and Hyndman 2017; Shang and Haberman 2018; Gao et al. 2019; Shang 2019). With this configuration and arranging the 47 prefectures together, we deal with a male, female, or total Japanese mortality data set of size 2007. Each prefecture does not have the same number of functions, and the range of observed years is also different. However, roughly, we have yearly mortality functions between 1975 and 2016.

In this case study, the poor availability of complete functions invalidates the possibility of resampling as done for the AEMET data set. Thus, we are only able to illustrate some empirical situations. Figures 6 and 7 present two reconstruction problems and the results obtained from the three methods. In Fig. 6, we reconstruct the shortest available curve, "Saitama-2007", that was only available in a very short interval of mortality rates for the youngest cohorts. Reg. Regression (Kraus 2015) and Opt. Operator (Kneip and Liebl 2020) methods produce smooth results (black and green, respectively). The depth-based method produces more spiky results in concordance with other available curves. The bottom panel presents the bubble plot illustrating the period of the envelope functions and the prefectures on the map. Figure 7 presents a case with the function "Tottori-2015" that is not observed in six intervals fragments (domain where only the red curve is visible).

4 Conclusion

This article introduces a non-parametric method to reconstruct samples of incomplete functional data. Our proposal relies on the concept of depth for POFD to select a subset of sample curves defined to share shape and magnitude with the observed

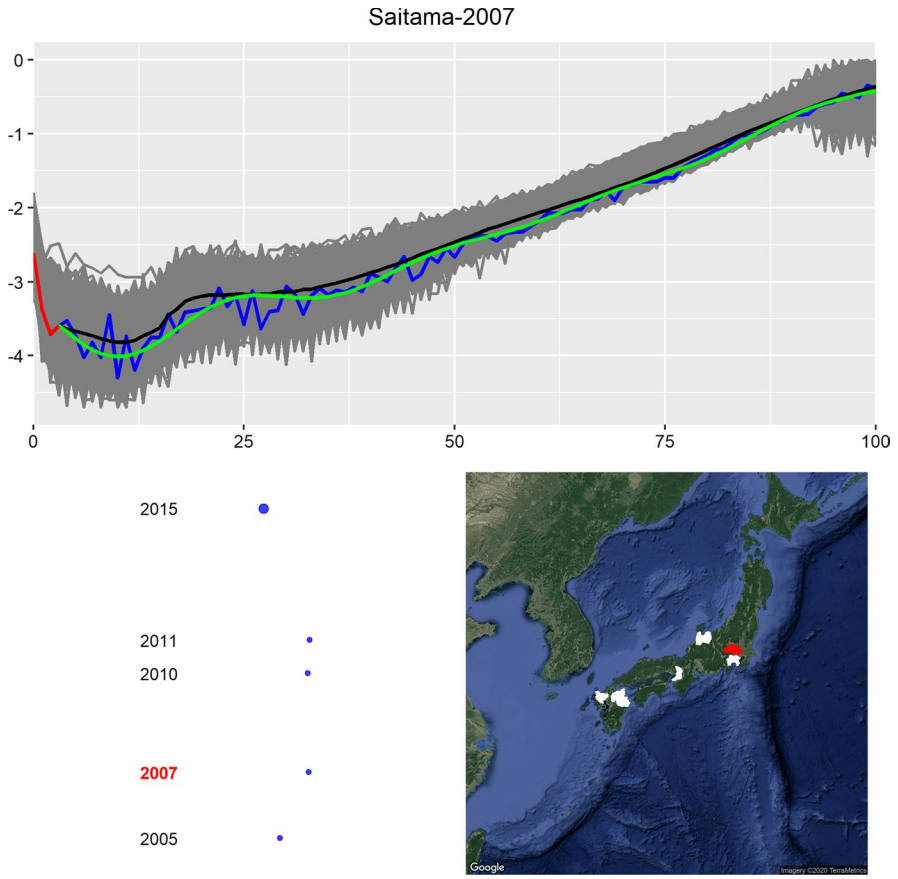

Fig. 6 Reconstruction of the most poorly observed function of the sample, "Saitama-2007". The top panel presents the reconstruction given by Reg. Regression method by Kraus (2015) (black), Opt. Operator method by Kneip and Liebl (2020) (green) and the depth-based method (blue). The bottom panel shows the year of the most important functions of the envelope, and the maps show the corresponding prefectures, in red the one to be reconstructed

part of the curve to predict. This subset of sample curves is termed envelope, and we use it to propose a point reconstruction method based on a weighted average. These weights only depend on a parameter we set by minimizing the MSPE where the curve to predict is observed.

We compare the new method's performance with other alternatives in the literature. Our simulation exercises consider simulated and empirical data as well as various random procedures to generate incomplete data scenarios. The available reconstruction methods seem unbeatable when covariance can be efficiently estimated in our settings. Gaussian processes exemplify these favorable circumstances with stationary covariance functions and other more complex covariance regimes, including a considerable proportion of completely observed curves. In contrast, our method outperforms when the covariance can not be properly estimated due to a richer

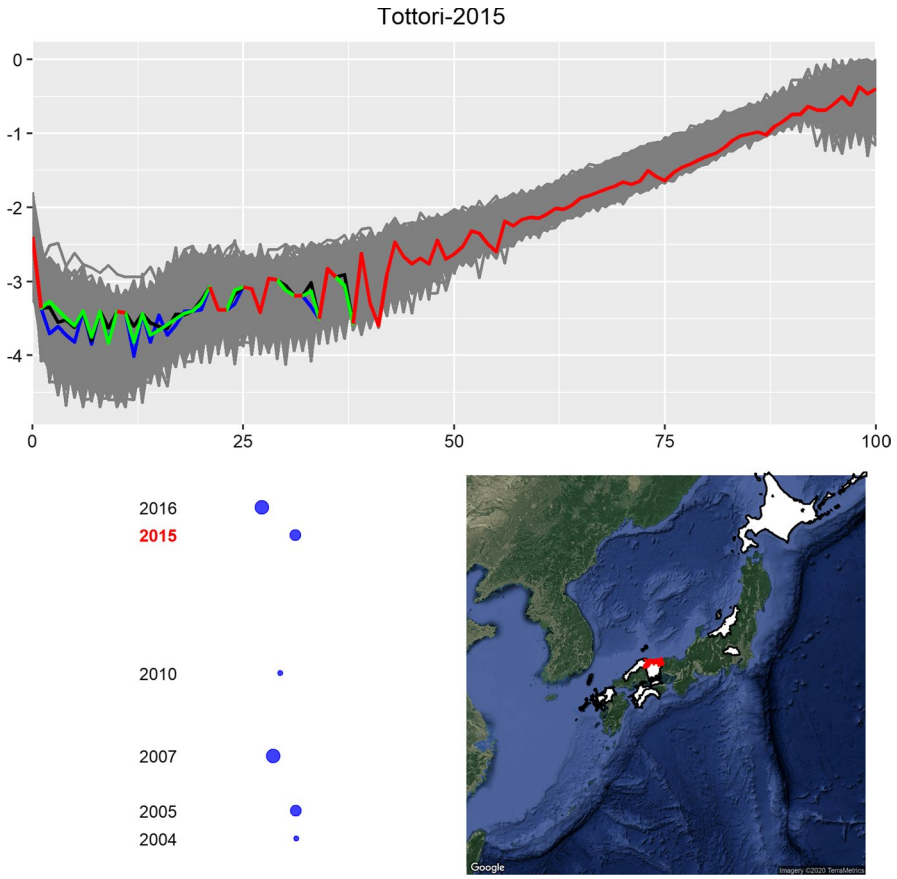

Fig. 7 Holdout partially observed sample function observed in six fragments, Tottori-2015. The top panel presents the reconstruction given by Reg. Regression method by Kraus (2015) (black), Opt. Operator method by Kneip and Liebl (2020) (green) and the depth-based method (blue). The bottom panel shows the year of the most important functions of the envelope, and the right map shows its prefectures

covariance structure and highly scarce data settings. To show that, we test the methods under severe incomplete data settings and introduce more complex covariance structures. We decrease the number of completely observed functions up to zero and consider empirical data with complex covariance structures, such as yearly age-specific mortality and temperature data. Finally, our simulation exercises show that our proposal can provide a reasonable reconstruction output when every function is partially observed or, in other words, when there are no complete functions in the sample.

The depth-based method requires the Missing-Completely-at-Random assumption, which is standard in the literature. This assumption implies that the partially observed functions cover densely the reconstruction domain and that the observability process is not conditional to external information. Future research could allow for specific relationships between the functional process and generate partial observability. In addition, the depth-based algorithm requires a notion of proximity between

POFD that we fill with a L_2 distance between the observed segments. Developments along this line would also be valuable for our proposal.

In summary, this article provides an alternative data-driven and model-free method to reconstruct POFD preferable under challenging scenarios. Last but not least, we believe that the interpretability of the results might help provide different insides into the data under analysis.

Appendix 1: Additional simulation results

Appendix 1.1: Gaussian processes

The random functions are generated as explained in Sect. 3.1. Then, the Missing-Completely-at-Random procedures are applied to generate partially observed functional data (X_i, O_i). The simulation results are collated in Table 4, for Random Intervals, and in Table 5, for Random Points.

Appendix 1.2: Karhunen-Loève processes

The random functions are generated by Karunhen-Loève expansions following the data generation processes in Kneip and Liebl (2020). Each function is generated as: $X_i(t) = \mu(t) + \sum_{k=1}^{50} \xi_{ik,1} cos(k\pi t) + \xi_{ik,1} sin(k\pi t)$ where $\mu(t) = t + \sin(2\pi t)$, $\xi_{ik,1} = 50\sqrt{\exp(-(k-1)^2)} Z_{i,1}$ and $\xi_{ik,2} = 50\sqrt{\exp(-k^2)} Z_{i,2}$ with $Z_{i,1}, Z_{i,2} \sim N(0,1)$. Then, the Missing-Completely-at-Random procedures are applied to generate POFD (X_i, O_i). The simulation results are collated in Table 6, for Random Intervals, and in Table 7, for Random Points.

Appendix 1.3: Karhunen-Loève processes with multiple populations

We follow the data generation procedure of Sect. A.2 but we create ten different populations by modifying the mean function $\mu(t)$ as follows: $\mu_1(t) = \sin(10\pi t)$, $\mu_2(t) = -\sin(10\pi t)$, $\mu_3(t) = \cos(10\pi t)$, $\mu_4(t) = -\cos(10\pi t)$, $\mu_5(t) = 5t + \sin(10\pi t + 0.5)$, $\mu_6(t) = 5t - \sin(10\pi t + 0.5)$ $\mu_6(t) = 5t - \sin(10\pi t + 0.5)$, $\mu_7(t) = 5t + \cos(10\pi t + 0.5)$, $\mu_8(t) = 5t - \cos(10\pi t + 0.5)$, $\mu_9(t) = 2t + \cos(3\pi t - 0.5)$ and $\mu_1 0(t) = -5t - \cos(4\pi t \sin(4\pi t)$. These populations are equally probable. The Missing-Completely-at-Random procedures are applied to generate partially observed functional data (X_i, O_i). The simulation results are collated in Table 8, for Random Intervals, and in Table 9, for Random Points.

Appendix 1.4: Computational time performance

See Table 10.

Table 4 Median values of the MSPE after 100 pseudo-random replicates

Intervals	Method	$c = 75$			$c = 50$			$c = 25$			$c = 0$		
		$p = 25$	50	75	25	50	75	25	50	75	25	50	75
$m = 1$	Depth-based	0.5045	0.4049	0.2938	0.5352	0.4133	0.3017	0.5755	0.4345	0.3184	**1.5962**	**0.5231**	**0.4352**
		(0.543)	(0.4671)	(0.3351)	(0.5512)	(0.4714)	(0.3592)	(0.604)	(0.497)	(0.3984)	(**1.7852**)	(**0.6501**)	(**0.5754**)
	Opt. operator	0.5445	0.4311	0.3378	0.528	0.4611	0.357	0.5411	0.4833	0.3853	–	–	–
		(0.5626)	(0.4797)	(0.3785)	(0.5664)	(0.4925)	(0.3791)	(0.6041)	(0.5414)	(0.4519)	–	–	–
	Reg. regression	**0.4198**	**0.3273**	**0.212**	**0.4207**	**0.3371**	**0.2143**	**0.4397**	**0.3566**	**0.2389**	–	–	–
		(**0.4383**)	(**0.3783**)	(**0.2424**)	(**0.4583**)	(**0.3887**)	(**0.2717**)	(**0.495**)	(**0.4408**)	(**0.3398**)	–	–	–
$m = 2$	Depth-based	0.4582	0.3348	0.2312	0.503	0.3433	0.2345	0.5223	0.3587	0.2484	**0.5696**	**0.4148**	**0.3219**
		(0.4592)	(0.3719)	(0.2786)	(0.4995)	(0.3807)	(0.2926)	(0.546)	(0.4107)	(0.3133)	(**0.7296**)	(**0.5165**)	(**0.4373**)
	Opt. operator	0.6553	0.4723	0.296	0.6544	0.4608	0.3095	0.6362	0.45	0.3446	–	–	–
		(0.6619)	(0.4746)	(0.2957)	(0.6231)	(0.4412)	(0.3142)	(0.5891)	(0.4663)	(0.3773)	–	–	–
	Reg. regression	**0.2995**	**0.2345**	**0.1407**	**0.3096**	**0.238**	**0.1484**	**0.3271**	**0.2602**	**0.1677**	–	–	–
		(**0.3284**)	(**0.2806**)	(**0.2026**)	(**0.3607**)	(**0.2893**)	(**0.2089**)	(**0.4066**)	(**0.3594**)	(**0.2633**)	–	–	–
$m = 4$	Depth-based	0.4098	0.2416	0.1793	0.438	0.2497	0.1858	0.46	0.2652	0.1922	**0.4913**	**0.3129**	**0.2369**
		(0.3682)	(0.2875)	(0.2377)	(0.4158)	(0.3027)	(0.236)	(0.4748)	(0.3253)	(0.2456)	(**0.5684**)	(**0.3894**)	(**0.3203**)
	Opt. operator	0.5763	0.3623	0.2404	0.5737	0.3539	0.2064	0.522	0.358	0.2488	–	–	–
		(0.5382)	(0.3461)	(0.2305)	(0.4903)	(0.3334)	(0.2687)	(0.4631)	(0.381)	(0.2918)	–	–	–
	Reg. regression	**0.197**	**0.1509**	**0.0942**	**0.209**	**0.162**	**0.1016**	**0.2271**	**0.184**	**0.1158**	–	–	–
		(**0.2286**)	(**0.1919**)	(**0.1454**)	(**0.2561**)	(**0.224**)	(**0.1609**)	(**0.3147**)	(**0.2803**)	(**0.2067**)	–	–	–

Each replicate is composed of 500 and 100 (results between parenthesis) functional observations. The partially observed samples are obtained by restricting each function to m intervals of total length $p\%$ of the domain (Random Intervals). The smallest error is bolded for each combination of c, m, and p

Table 5 Median values of MSPE over 100 pseudo-random replicates

Method	$c = 75$			$c = 50$			$c = 25$			$c = 0$		
	$p = 25$	50	75	25	50	75	25	50	75	25	50	75
Depth-based	0.1336	0.1171	0.1123	0.1512	0.1214	0.115	0.1738	0.1275	0.1164	**0.2163**	**0.1368**	**0.1193**
	(0.1746)	(0.1636)	(0.1597)	(0.1927)	(0.1711)	(0.1571)	(0.2292)	(0.1835)	(0.1650)	(**0.3145**)	(**0.2005**)	(**0.1714**)
Opt. operator	0.1527	0.1639	0.2351	0.1408	0.1547	0.2403	0.1515	0.1495	0.2201	–	–	–
	(0.1699)	(0.1731)	(0.2471)	(0.165)	(0.1783))	(0.2489)	(0.1871)	(0.1718)	(0.2503)	–	–	–
Reg. regression	**0.0436**	**0.0366**	**0.0353**	**0.0541**	**0.0507**	**0.0428**	**0.0769**	**0.0656**	**0.0498**	–	–	–
	(**0.0846**)	(**0.0815**)	(**0.0773**)	(**0.1095**)	(**0.1046**)	(**0.0892**)	(**0.1917**)	(**0.1782**)	(**0.1418**)	–	–	–

Each replicate composes 500 and 100 curves (results between parenthesis). A dash (-) represents that the method cannot produce any reconstruction. The partially observed samples are obtained by observing p% of the total discrete realization points (Random Points). The smallest error is bolded for each combination of c and p

Table 6 Median values of the MSPE after 100 pseudo-random replicates

Intervals	Method	$c = 75$			$c = 50$			$c = 25$			$c = 0$		
		$p = 25$	50	75	25	50	75	25	50	75	25	50	75
$m = 1$	Depth-based	0.1929	0.1202	0.0431	0.2226	0.1257	0.049	0.2655	0.1414	0.0637	**0.6216**	**0.2229**	**0.1734**
		(0.2386)	(0.1764)	(0.0923)	(0.2639)	(0.2031)	(0.1053)	(0.3213)	(0.2318)	(0.1421)	(**0.6985**)	(**0.4293**)	(**0.312**)
	Opt. operator	0.1615	0.1045	0.0491	0.1642	0.1052	0.0473	0.1699	0.1204	0.0522	–	–	–
		(0.1754)	(0.1209)	(0.0536)	(0.1805)	(0.134)	(0.0571)	(**0.2087**)	(**0.1571**)	(**0.0701**)	–	–	–
	Reg. regression	**0.0662**	**0.0337**	**0.0098**	**0.101**	**0.0477**	**0.0148**	**0.1388**	**0.0709**	**0.0282**	–	–	–
		(**0.1138**)	(**0.0652**)	(**0.0376**)	(**0.1428**)	(**0.0951**)	(**0.0487**)	(0.2117)	(0.1601)	(0.0823)	–	–	–
$m = 2$	Depth-based	0.153	0.0622	0.024	0.2189	0.0768	0.0289	0.2865	0.1012	0.036	**0.4121**	**0.1742**	**0.1203**
		(0.166)	(0.1063)	(0.0563)	(0.2226)	(0.1295)	(0.0673)	(0.3142)	(0.1619)	(0.0841)	(**0.5581**)	(**0.2764**)	•
	Opt. operator	0.3156	0.0861	0.0176	0.3232	0.0898	0.0167	0.3308	0.0979	0.0203	–	–	–
		(0.2935)	(0.0983)	(0.0164)	(0.3054)	(0.1094)	(**0.0257**)	(0.3488)	(0.1296)	(**0.0332**)	–	–	–
	Reg. regression	**0.0133**	**0.0107**	**0.0036**	**0.0293**	**0.018**	**0.0065**	**0.0561**	**0.0276**	**0.0118**	–	–	–
		(**0.0256**)	(**0.0307**)	(**0.0122**)	(**0.0572**)	(**0.0465**)	(0.0261)	(**0.1331**)	(**0.0842**)	(0.0456)	–	–	–
$m = 4$	Depth-based	0.1486	0.0275	0.0147	0.2582	0.036	0.0172	0.3278	0.05	0.0215	**0.4079**	**0.1118**	**0.0839**
		(0.1209)	(0.0567)	(0.0376)	(0.1903)	(0.0707	(0.0441)	(0.2838)	(0.0933)	(0.0565)	(**0.4778**)	(**0.1775**)	(**0.1266**)
	Opt. operator	0.1063	0.0289	**0.002**	0.1113	0.0297	**0.0028**	0.1149	**0.0317**	**0.0048**	–	–	–
		(0.1174)	(0.0315)	(**0.0047**)	(0.1284)	(0.0358)	(**0.0078**)	(0.1536)	(**0.0511**)	(**0.0141**)	–	–	–
	Reg. regression	**0.0028**	**0.0043**	0.0026	**0.0068**	**0.0076**	0.0044	**0.0172**	0.0139	0.0083	–	–	–
		(**0.0066**)	(**0.016**)	(0.0105)	(**0.0197**)	(**0.028**)	(0.0187)	(**0.061**)	(0.0531)	(0.0336)	–	–	–

Each replicate is composed of 500 and 100 (results between parenthesis) functional observations. The partially observed samples are obtained by restricting each function to m intervals of total length $p\%$ of the domain (Random Intervals). The smallest error is bolded for each combination of c and p

Table 7 Median values of MSPE over 100 pseudo-random replicates

Method	$c = 75$			$c = 50$			$c = 25$			$c = 0$		
	$p = 25$	50	75	25	50	75	25	50	75	25	50	75
Depth-based	0.0694	0.0117	0.0104	0.1484	0.015	0.0112	0.2302	0.0184	0.0123	**0.3181**	**0.0253**	**0.0135**
	(0.0603)	(0.0327)	(0.0 282)	(0.1004)	(0.0395)	(0.0308)	(0.1684)	(0.0468)	(0.0345)	(**0.3191**)	(**0.0653**)	(**0.0377**)
Opt. operator	0.0113	0.0011	0.0579	0.0113	**0.0013**	0.0622	0.0134	**0.0019**	0.0578	–	–	–
	(0.0116)	(**0.0022**)	(0.0599)	(0.0133)	(**0.003**)	(0.0583)	(**0.0171**)	(**0.0095**)	(0.0556)	–	–	–
Reg. regression	**0.001**	**0.0028**	**0.002**	**0.0026**	0.0055	**0.0034**	**0.0077**	0.0103	**0.0052**	–	–	–
	(**0.0033**)	(0.0124)	(**0.0091**)	(**0.0097**)	(0.0226)	(**0.0141**)	(0.0361)	(0.0406)	(**0.0212**)	–	–	–

Each replicate is composed of 500 and 100 (results between parenthesis) functional observations. A dash (-) represents that the method cannot produce any reconstruction. The partially observed samples are obtained by observing $p\%$ of the total discrete realization points (Random Points). The smallest error is bolded for each combination of c and p

Table 8 Median values of the MSPE after 100 pseudo-random replicates

Intervals	Method	$c = 75$			$c = 50$			$c = 25$			$c = 0$		
		$p = 25$	50	75	25	50	75	25	50	75	25	50	75
$m = 1$	Depth-based	**0.3669**	0.2472	**0.138**	**0.4702**	0.3038	**0.1627**	0.6666	0.3834	**0.2117**	**2.123**	**0.754**	**0.4761**
		(0.622)	(0.4977)	(**0.3387**)	(0.8087)	(0.5958)	(**0.4179**)	(1.1973)	(**0.8202**)	(0.5777)	(**2.5212**)	(**1.6303**)	(**1.0853**)
	Opt. operator	1.4395	1.0841	0.7514	1.4402	1.0954	0.7843	1.4692	1.0979	0.8705	-	-	-
		(1.4899)	(1.1639)	(0.8117)	(1.5149)	(1.1283)	(0.8572)	(1.5931)	(1.2441)	(0.9856)	-	-	-
	Reg. regression	0.4138	**0.1435**	0.1818	0.5261	**0.1832**	0.2046	**0.6508**	**0.2887**	0.262	-	-	-
		(**0.548**)	(**0.4628**)	(0.3859)	(**0.7899**)	(**0.5073**)	(0.4209)	(**1.0905**)	(0.827)	(**0.572**)	-	-	-
$m = 2$	Depth-based	0.4422	0.1577	0.086	0.7784	0.1993	0.1041	1.1877	0.2703	0.1375	**1.9447**	**0.5022**	**0.2892**
		(0.6077)	(0.3711)	(0.24)	(0.957)	(0.4634)	(0.2997)	(1.4361)	(0.6541)	(0.4068)	(**2.6944**)	(**1.1145**)	(**0.636**)
	Opt. operator	1.8112	1.0331	0.711	1.8903	1.0468	0.7224	1.8395	1.0563	0.7316	-	-	-
		(1.734)	(1.038)	(0.7546)	(1.838)	(1.0817)	(0.7612)	(1.7209)	(1.1341)	(0.8135)	-	-	-
	Reg. regression	**0.1494**	**0.0788**	**0.0672**	**0.2199**	**0.0985**	**0.0803**	**0.3604**	**0.1713**	**0.1174**	-	-	-
		(**0.2262**)	(**0.2751**)	(**0.187**)	(**0.4511**)	(**0.3506**)	(**0.2441**)	(**0.7987**)	(**0.581**)	(**0.3713**)	-	-	-
$m = 4$	Depth-based	0.62	0.0875	0.0623	1.1437	0.1176	**0.0717**	1.5722	0.1626	**0.0923**	**1.9824**	**0.2896**	**0.1977**
		(0.6087)	(0.2636)	(0.2024)	(1.0035)	(0.3453)	(0.2262)	(1.5562)	(**0.46**	(0.303)	(**2.3563**)	(**0.694**)	(**0.3958**)
	Opt. operator	1.3571	0.8519	0.6423	1.3435	0.8497	0.6526	1.3426	0.8618	0.6579	-	-	-
		(1.3645)	(0.8743)	(0.6668)	(1.3573)	(0.8857)	(0.6769)	(1.3708)	(0.9156)	(0.6853)	-	-	-
	Reg. regression	**0.042**	**0.0453**	**0.0571**	**0.0854**	**0.0498**	0.0748	**0.1848**	**0.1092**	0.1031	-	-	-
		(**0.0926**)	(**0.2169**)	(**0.1609**)	(**0.2293**)	(**0.2988**)	(**0.2177**)	(**0.6179**)	(0.4715)	(**0.2904**)	-	-	-

Each replicate is composed of 500 and 100 (results between parenthesis) functional observations. The partially observed samples are obtained by restricting each function to m intervals of total length $p\%$ of the domain (Random Intervals). The smallest error is bolded for each combination of c, m, and p

Table 9 Median values of MSPE over 100 pseudo-random replicates

Method	$c = 75$			$c = 50$			$c = 25$			$c = 0$		
	$p = 25$	50	75	25	50	75	25	50	75	25	50	75
Depth-based	0.3838	0.0532	0.046	0.8209	0.0668	**0.0513**	1.2769	0.0 845	**0.057**	**1.6502**	**0.1141**	**0.0636**
	(0.3767)	(0.1985)	(0.1691)	(0.6992)	(**0.2458**)	(**0.1825**)	(1.0884)	(**0.3051**)	(**0.1994**)	(**1.7077**)	(**0.3929**)	(**0.2216**)
Opt. operator	0.7089	0.5843	0.6526	0.6956	0.5846	0.6518	0.6907	0.5869	0.653	–	–	–
	(0.6919)	(0.6006)	(0.6537)	(0.7189)	(0.593)	(0.6526)	(0.7273)	(0.601)	(0.6607)	–	–	–
Reg. regression	**0.0159**	**0.0235**	**0.0521**	**0.0377**	**0.0294**	0.0674	**0.0963**	**0.0613**	0.089	–	–	–
	(**0.0495**)	(**0.184**)	(**0.1407**)	(**0.127**)	(0.2699)	(0.1952)	(**0.3716**	(0.4085)	(0.2486)	–	–	–

Each replicate is composed of 500 and 100 (results between parenthesis) functional observations. A dash (-) represents that the method cannot produce any reconstruction. The partially observed samples are obtained by observing $p\%$ of the total discrete realization points (Random Points). The smallest error is bolded for each combination of c and p

Table 10 Computational time performance

Method	Gaussian	Karhunen-Loève	Karhunen-Loève Multi.
Depth-based	598.682	365.488	349.227
	(20.068)	(9.045)	(9.374)
Depth-based parallel (8 cores)	**74.835**	**45.686**	**43.653**
	(**2.508**)	(1.130)	(1.171)
Opt. operator	162.329	49.961	51.779
	(22.260)	(3.2647)	(4.015)
Reg. regression	94.480	**7.0162**	**6.899**
	(9.664)	**(0.450)**	**(0.517)**

Mean in seconds over all the replicates (1800) for each data generation process. The first row of each method is the time for a sample size of 500 and the second row for a sample size of 100. The smallest computational time is bolded

Acknowledgements The authors are grateful for insightful comments and suggestions from the Associate Editor and a reviewer. Antonio Elías was supported by the Ministerio de Educación, Cultura y Deporte under grant FPU15/00625 and the research stay grant EST17/00841. Antonio Elías and Raúl Jiménez were partially supported by the Spanish Ministerio de Economía y Competitividad under grants ECO2015-66593-P and PID2019-109196GB-I00. Part of this article was conducted during a stay at the Australian National University. Antonio Elías is grateful to Han Lin Shang for his hospitality and insightful and constructive discussions. The authors thankfully acknowledge the computer resources, technical expertise, and assistance provided by the Supercomputing and Bioinformatics center of the University of Málaga.

Author Contributions (Following CRediT author statement) AE: Writing - original draft. writing - review and editing. Conceptualization. methodology. Software. Data curation. Visualization. RJ Writing - original draft. Writing - review and editing. Conceptualization. Methodology. Supervision. HLS writing - review and editing. Conceptualization. Methodology. Supervision.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Funding for open access charge: Universidad de Málaga / CBUA. Not applicable.

Availability of data and material The required links to download the data are in the main text.

Code availability The available code is referenced in the main text.

Declarations

Competing interests Not applicable.

References

Claeskens G, Hubert M, Slaets L, Vakili K (2014) Multivariate functional halfspace depth. J Am Stat Assoc Theory Methods 109(505):411–423

D'Amato V, Piscopo G, Russolillo M (2011) The mortality of the Italian population: smoothing techniques on the Lee-Carter model. Ann Appl Stat 5(2A):705–724

Delaigle A, Hall P (2013) Classification using censored functional data. J Am Stat Assoc Theory Methods 108(504):1269–1283

Delaigle A, Hall P (2016) Approximating fragmented functional data by segments of markov chains. Biometrika 103(4):779–799

Elías A, Jiménez R, Paganoni A, Sangalli L (2021) fdaPOIFD: partially observed integrated functional depth. R package version 1.0.0. https://CRAN.R-project.org/package=fdaPOIFD

Elías A, Jiménez R, Paganoni AM, Sangalli LM (2022a) Integrated depth for partially observed functional data. J Computat Gr Stat.

Elías A, Jiménez R, Shang HL (2022) On projection methods for functional time series forecasting. J Multivar Anal 189:104890

Febrero-Bande M, Oviedo de la Fuente M (2012) Statistical computing in functional data analysis: the R package fda.usc. J Stat Softw, 51(4):1–28

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

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

Gao Y, Shang HL, Yang Y (2019) High-dimensional functional time series forecasting: an application to age-specific mortality rates. J Multivar Anal 170:232–243

García-Portugués E, González-Manteiga W, Febrero-Bande M (2014) A goodness-of-fit test for the functional linear model with scalar response. J Comput Graph Stat 23(3):761–778

Goldberg Y, Ritov Y, Mandelbaum A (2014) Predicting the continuation of a function with applications to call center data. J Stat Plann Inference 147:53–65

Hubert M, Rousseeuw PJ, Vanden Branden K (2005) Robpca: a new approach to robust principal component analysis. Technometrics 47(1):64–79

Human Mortality Database (2022) University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany). www.mortality.org. Accessed July 21, 2021

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

James GM, Hastie TJ (2001) Functional linear discriminant analysis for irregularly sampled curves. J R Stat Soc Ser B (Stat Methodol) 63(3):533–550

Kneip A, Liebl D (2020) On the optimal reconstruction of partially observed functional data. Ann Stat 48(3):1692–1717

Kraus D (2015) Components and completion of partially observed functional data. J R Stat Soc Ser B (Stat Methodol) 77(4):777–801

Kraus D, Stefanucci M (2020) Ridge reconstruction of partially observed functional data is asymptotically optimal. Stat Probab Lett 165:108813

Li J, Cuesta-Albertos JA, Liu RY (2012) *DD*-classifier: nonparametric classification procedure based on *DD*-plot. J Am Stat Assoc Theory Methods 107(498):737–753

Liebl D, Rameseder S (2019) Partially observed functional data: the case of systematically missing parts. Comput Stat Data Anal 131:104–115

Liu RY (1990) On a notion of data depth based on random simplices. Ann Stat 18(1):405–414

Liu RY, Parelius JM, Singh K (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference. Ann Stat 27(3):783–840

López-Pintado S, Romo J (2009) On the concept of depth for functional data. J Am Stat Assoc Theory Methods 104(486):718–734

Mozharovskyi P, Josse J, Husson F (2020) Nonparametric imputation by data depth. J Am Stat Assoc Theory Methods 115(529):241–253

Nagy S, Gijbels I, Omelka M, Hlubinka D (2016) Integrated depth for functional data: statistical properties and consistency. ESAIM, Probability and Statistics, p 20

Ramsay J, Silverman B (2005) Functional data analysis. Springer, New York, 2nd edition

Rousseeuw PJ, Ruts I, Tukey JW (1999) The bagplot: a bivariate boxplot. Am Stat 53(4):382–387

Sangalli LM, Secchi P, Vantini S (2014) AneuRisk65: a dataset of three-dimensional cerebral vascular geometries. Electron J Stat 8(2):1879–1890

Sangalli LM, Secchi P, Vantini S, Veneziani A (2009) A case study in exploratory functional data analysis: geometrical features of the internal carotid artery. J Am Stat Assoc Appl Case Stud 104(485):37–48

Serfling RJ (2006) Multivariate symmetry and asymmetry. In: Kotz S, Read CB, Balakrishnan N, Vidakovic B, Johnson NL (eds) Encyclopedia of statistical sciences, volume 8, pp 5338–5345. Wiley-Interscience, Hoboken, New Jersey, 2nd edition

Shang HL (2019) Visualizing rate of change: an application to age-specific fertility rates. J R Stat Soc A Stat Soc 182(1):249–262

Shang HL, Haberman S (2018) Model confidence sets and forecast combination: an application to age-specific mortality. Genus 74(1):19

Shang HL, Hyndman RJ (2017) Grouped functional time series forecasting: an application to age-specific mortality rates. J Comput Graph Stat 26(2):330–343

Yao F, Muller HG, Wang JL (2005) Functional data analysis for sparse longitudinal data. J Am Stat Assoc Theory Methods 100(470):577–590

Zuo Y, Serfling R (2000) General notions of statistical depth function. Ann Stat 28(2):461–482