Journal de la Société Française de Statistique
Vol. 155 No. 4 (2014)

Analysing large datasets of functional data: a survey sampling point of view

Titre: Analyse statistique de grandes bases de données fonctionnelles : le point de vue de sondeurs

Pauline Lardin-Puech[1], Hervé Cardot[2] and Camelia Goga[2]

Abstract: At the age of Big Data, it is now common to have to deal with very large datasets of phenomena that evolve over time. When the aim is to estimate simple quantities such as the mean or the median trajectory, as well as the main modes of variation of the data, captured through a principal components analysis, survey sampling techniques may be employed successfully. They can offer an interesting trade off between size of the data and accuracy of estimators. This paper makes a review of survey sampling approaches recently developed to deal with large datasets of functional data. We present different sampling techniques that can be employed to build confidence bands and improve, with the help of auxiliary information, the accurary of estimators compared to simple random sampling without replacement. These procedures are illustrated on a dataset of electricity load curves measured every half-hour over a period of one week.

Résumé : A l'ère des données massives, il n'est plus inhabituel d'avoir à gérer de très grandes bases de données de phénomènes temporels. Quand l'objectif est d'estimer des indicateurs simples tels que la trajectoire moyenne ou médiane ou bien encore les principaux modes de variation autour de la moyenne, capturés par l'intermédiaire d'une analyse en composantes principales, les techniques de sondage sont des approches intéressantes. Elles offrent en effet un bon compromis entre taille des données à traiter et précision de l'estimation. Ce travail présente une revue des approches de sondage qui ont été developpées ces dernières années pour analyser de grandes bases de données fonctionnelles. L'accent est mis sur les manières de prendre en compte l'information auxiliaire en vue d'améliorer l'estimation en comparaison avec le sondage aléatoire simple sans remise et sur la construction de bandes de confiance. Ces techniques sont illustrées sur un jeu de données de courbes de charge électrique mesurées chaque demi-heure pendant une semaine.

Keywords: Big Data, Confidence bands, Horvitz-Thompson estimator, Model-assisted estimation, Unequal probability sampling designs, Variance estimation
Mots-clés : Bandes de confiance, Données massives, Estimateur de Horvitz-Thompson, Estimateurs assistés par un modèle, Estimation de la variance, Plans à probabilités inégales
AMS 2000 subject classifications: 62D05, 60F17, 62G05

1. Introduction

Steve Lohr wrote in the New York Times (see Lohr (2012)) that we are now at "the Age of Big Data", with "countless digital sensors worldwide in industrial equipment, automobiles, electrical meters and shipping crates". An example of such large data of phenomena that evolve over time is given by electricity load curves measured for households and companies thanks to new smart electricity meters. Such data have been studied in Pauline Lardin's thesis Lardin (2012),

[1] EDF R&D - La Poste.
and E-mail: pauline.puech@laposte.fr
[2] Université de Bourgogne. Institut de Mathématiques de Bourgogne, UMR CNRS 5584. 9 Av. Alain Savary, 21078 Dijon Cedex. France.
E-mail: herve.cardot@u-bourgogne.fr and E-mail: camelia.goga@u-bourgogne.fr

 ISSN: 2102-6238

with the financial support of Electricité de France (EDF), the major French electricity company. In the presence of technical and budgetary constraints due to limited bandpass or storage cost of huge databases, the analysis of the whole set may be impossible or very difficult. In Chiky (2009), it is shown that if we are only interested in simple indicators, such as total or mean trajectories, even very simple survey sampling techniques, such as simple random sampling without replacement, are attractive alternatives to signal compression techniques since they permit to obtain precise estimates at a reasonable cost. The aim of this work was to evaluate how survey sampling techniques could be useful and how they can be adapted to deal with observations that are functions of time when one aims at estimating functional parameters of interest such as the mean load curves. Other references on this topic are Chiky et al. (2008) and Cardot and Josserand (2011).

We consider a test population of $N = 18902$ French companies whose electricity consumption has been measured every half an hour over a period of one week. A sample of 20 load curves extracted from this dataset is drawn in Figure 1 as well as the mean and the median profiles.

The discretization scheme is very fine so that the statistical units can be considered as functions of time. We can use the tools of functional data analysis to describe the data and build statistical models. Even if some of these tools have been first proposed in the 1970s in Deville (1974) and Dauxois and Pousse (1976), these methods only began to spread twenty years ago with the increase of computer performances as well as storage capacities. The reader may refer to Ramsay and Silverman (2005), Ferraty and Vieu (2006) and Ferraty and Romain (2011) for an overview of the different techniques developed in the statistical literature in functional data analysis as well as examples of application.

This work aims at giving a review of some recent works combining survey sampling techniques and functional data analysis. We focus on the estimation of simple quantities such as mean, principal components or medians and explain how auxiliary information can be taken into account in order to improve the accuracy of the estimators compared to simple random sampling without replacement. Under asymptotic normality assumptions, we also present how it is possible to build confidence bands when we have at hand a consistent estimator of the variance.

The paper is structured as follows. Notations and functional parameters of interest are given in Section 2. Estimators for the functional parameters are suggested in Section 3 and their asymptotic properties are studied in Section 4. Section 5 deals with an application on electricity load curves for stratified and πps sampling and we suggest in Section 6 an estimator that takes into account the auxiliary information by considering a functional linear model. Finally, Section 7 contains some concluding remarks.

2. Notations and parameters of interest

We consider a finite population U whose size N is not necessarily known. We suppose that for each unit k from the population U, we can observe a deterministic function of time $Y_k = (Y_k(t))_{t\in[0,\mathscr{T}]}$ that belongs to some space of functions. Depending on the objective, this space will be either the space of continuous functions $C[0,\mathscr{T}]$ endowed with the sup norm or the Hilbert space $L^2[0,\mathscr{T}]$, *i.e.* the space of square integrable functions defined on the closed interval $[0,\mathscr{T}]$, equipped with the inner product $< f,g >= \int_0^{\mathscr{T}} f(t)g(t)dt$ and the induced norm $||f|| = [\int_0^{\mathscr{T}} f^2(t)d(t)]^{1/2}$ for $f,g \in [0,\mathscr{T}]$.

 ISSN: 2102-6238

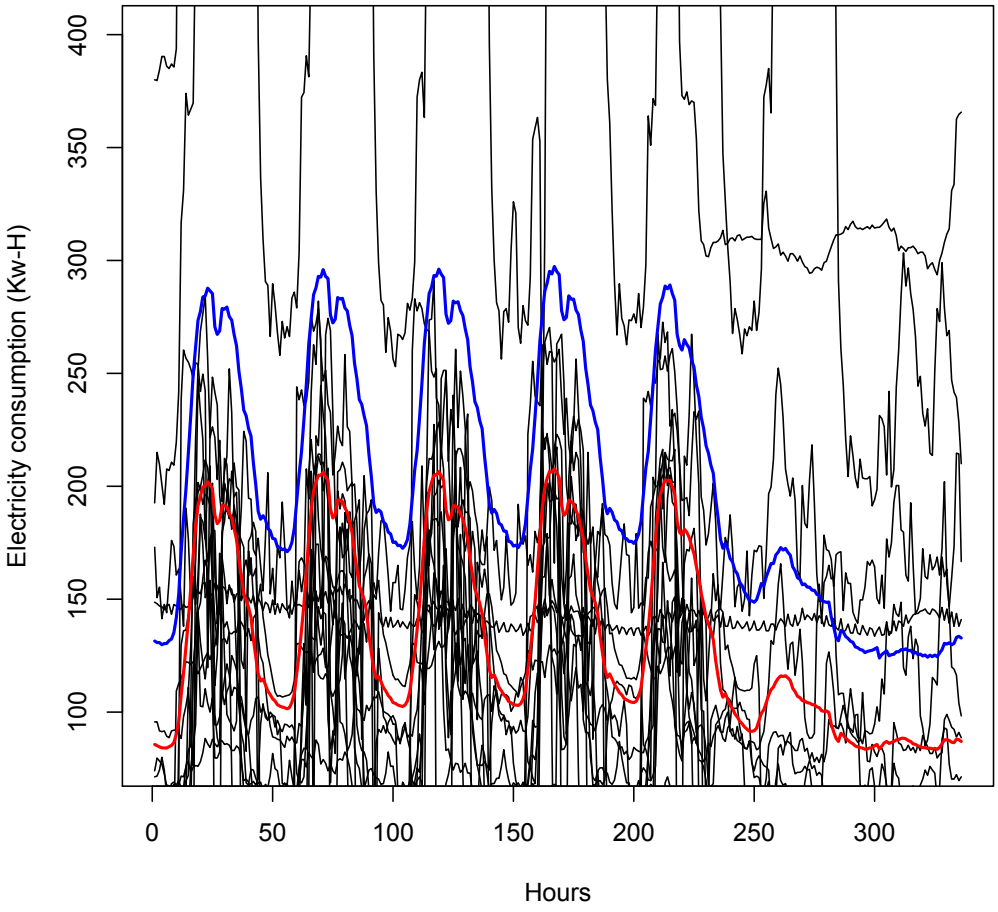


FIGURE 1. *A sample of 20 electricity consumption curves measured every half an hour over a period of one week. The mean consumption curve in the population is drawn in bold blue line and the median curve in red one.*

In this functional setting, the statistician may be interested in estimating classical parameters of interest such as the total or the mean curve and their definition and interpretation are obtained easily by analogy with the non-functional case. The situation is more complicated for other parameters of interest, such as quantiles. The median may be defined in several manners for multivariate or functional data. Moreover, new functional parameters may be of interest now.

When the aim is to build confidence bands, the natural setting will be the space $C[0, \mathscr{T}]$ since we want to produce a confidence interval that is uniform in t. When the aim is to estimate the principal components, it is required that the underlying functional space would be equipped with an inner product, so that the natural setting is to consider that Y_k are elements of $L^2[0, \mathscr{T}]$. The same functional space $L^2[0, \mathscr{T}]$ will be considered in the case of the median curve since the strict

 ISSN: 2102-6238

convexity of the norm $||\cdot||$ permits to obtain the uniqueness of this parameter.

We present below the functional parameters of interest that have been studied in a survey sampling setting. The simplest ones are the total curve:

$$t_Y = \sum_{k\in U} Y_k$$

and the mean trajectory:

$$\mu_N = \frac{1}{N}\sum_{k\in U} Y_k. \tag{1}$$

The value of t_Y or μ_N in a measurement point $t \in [0,\mathscr{T}]$ is obtained directly as $t_Y(t) = \sum_{k\in U} Y_k(t)$ and $\mu_N(t) = \frac{1}{N}\sum_{k\in U} Y_k(t)$, respectively.

For such high dimensional data, other useful statistical indicators are given by the principal components that can exhibit the main modes of variation of the data around the mean curve (see *e.g* Ramsay and Silverman (2005) and Cardot et al. (2010a)). To perform principal components analysis, it is first required to estimate the covariance function of the data at the population level. For r and t in $[0,\mathscr{T}]$, the covariance function $\gamma(r,t)$ between $(Y_k(r))_{k\in U}$ and $(Y_k(t))_{k\in U}$ is defined as follows:

$$\gamma(r,t) = \frac{1}{N}\sum_{k\in U}(Y_k(r)-\mu_N(r))(Y_k(t)-\mu_N(t)), \quad (r,t)\in[0,\mathscr{T}]\times[0,\mathscr{T}].$$

Then, the associated covariance operator Γ, which maps any function a in $L^2[0,\mathscr{T}]$ to Γa in $L^2[0,\mathscr{T}]$, is defined by

$$\Gamma a(r) = \int_0^{\mathscr{T}} \gamma(r,t)a(t)dt, \quad r\in[0,\mathscr{T}]. \tag{2}$$

The covariance operator has the following equivalent form:

$$\Gamma = \frac{1}{N}\sum_{k\in U}(Y_k-\mu_N)\otimes(Y_k-\mu_N), \tag{3}$$

where the tensor product of two elements a and b of $L^2[0,\mathscr{T}]$ is the rank one operator such that $a\otimes b(y) =< a,y > b$ for all $y\in L^2[0,\mathscr{T}]$. The eigenvalues of Γ are non negative and supposed to be sorted in decreasing order $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq 0$. They satisfy:

$$\Gamma v_j(t) = \lambda_j v_j(t), \quad j = 1,\ldots,N, \tag{4}$$

where the eigenfunctions $v_j, j = 1,\ldots,N$ can be chosen to form an orthonormal system in $L^2[0,\mathscr{T}]$, namely $< v_j, v_{j'} >= 1$ if $j = j'$ and zero otherwise.

With functional data, curves Y_k usually span a subspace whose dimension can be very large, at most N but with redundant information so that dimension reduction can be useful to analyze and describe the data. A classical tool for dimension reduction is principal components analysis (see Jolliffe (2002) or Ramsay and Silverman (2005)) which has been adapted in a finite population context by Cardot et al. (2010a) to get the best representation, in a quadratic sense, of the

 ISSN: 2102-6238

curves Y_k, in a subspace of dimension q much smaller than N. We consider the projection onto a q-dimensional subspace of $Y_k - \mu_N$ and we look for the subspace which can reproduce the best the variability of the data in the population. In order to achieve this, we look for the minimum of the following loss function:

$$R(q) = \frac{1}{N} \sum_{k=1}^{N} \|R_{qk}\|^2 \tag{5}$$

where

$$R_{qk}(t) = Y_k(t) - \mu_N(t) - \sum_{j=1}^{q} < Y_k - \mu_N, \phi_j > \phi_j(t), \quad t \in [0,T] \tag{6}$$

among all the orthonormal systems $\phi_1, \ldots, \phi_q$ in $L^2[0,\mathscr{T}]$ which span a q-dimensional subspace of $L^2[0,\mathscr{T}]$. We can remark that $\sum_{j=1}^{q} < Y_k - \mu_N, \phi_j > \phi_j$ is simply the orthogonal projection of $Y_k - \mu_N$ onto the subspace generated by $\phi_1, \ldots, \phi_q$. It can be shown (see Cardot et al. (2010a)) that the minimum of $R(q)$ is attained when considering the subspace generated by the q eigenfunctions $v_1, \ldots, v_q$ associated to the q largest eigenvalues of Γ and that, at the optimum, $R(q) = \sum_{j=q+1}^{N} \lambda_j$. Nevertheless, in practice we are not able to realize this decomposition since μ_N and Γ, as well as the eigenvalues λ_j and the eigenfunctions v_j, are computed from the whole population, as seen from equations (3) and (4). Instead, one can estimate μ_N, Γ, λ_j and v_j by drawing a sample as studied in Cardot et al. (2010a).

With high dimensional data, it is not uncommon to have outlying curves, such as consumers with very high levels of electricity consumption. In such a situation, it is advisable to consider indicators which are more robust to outlying data than the mean profile, and the median is one of them. However, the notion of median can not be generalized easily to multivariate or functional data because of the lack of a natural ordering. There are several definitions of the median and we present here the one used by Kemperman (1987) and Gervini (2008) for functional data. The reader is referred to Small (1990) for a review of different definitions of the median with multidimensional data. The median curve calculated from the elements $Y_1, \ldots, Y_N$ belonging to $L^2[0,\mathscr{T}]$ is defined by:

$$m_N = \text{argmin}_{y \in L^2[0,\mathscr{T}]} \sum_{k=1}^{N} \|Y_k - y\|. \tag{7}$$

If the points Y_k, for $k = 1, \ldots, N$, are not concentrated on a line, the median m_N exists and is unique (see Kemperman (1987)). For $Y_1, \ldots, Y_N \in \mathbb{R}^d$, m_N defined by the relation (7) arises as a natural generalization of the well-known characterization of the univariate median Koenker and Bassett (1978), $m_N = \arg\min_{y \in \mathbb{R}} \sum_{k=1}^{N} |Y_k - y|$. This indicator has been used for the first time at the beginning of the 20-th century. It was called the *spatial median* by Brown (1983) because, from a geometric point of view, the median is the point that minimizes the sum of distances to the points in the population. For example, Weber (1909) considered the following problem: a company wants to find the optimal location of its warehouse in order to serve the N customers with planar coordinates given by $Y_1, \ldots, Y_N$. The name of *L_1-median* was used by Small (1990) because the definition uses a L_1-criterion. Finally, Chaudhuri (1996) called it the *geometric median* because it

 ISSN: 2102-6238

may be seen as a particular case of the geometric quantiles whose definition uses the geometry of the data clouds by means of a direction and a magnitude.

The median defined in this way is a global indicator of the data in the sense that it takes into account all the measurement instants. Besides, it is a central indicator of the distribution of the data with nice robustness properties; see Ilmonen et al. (2012) for a recent review of the properties of the L_1-median.

It can also be shown that the median m_N is the unique solution of the following estimating equation (see Small (1990)),

$$\sum_{k=1}^{N} \frac{Y_k - y}{||Y_k - y||} = 0 \tag{8}$$

provided that we don't have $Y_k = m_N$ for any $k = 1, \ldots, N$.

The median defined by (7) or (8) may be computed by using fast iterative algorithms such as Weiszfeld's algorithm (see Weiszfeld (1937) and Vardi and Zhang (2000)) for multivariate data or gradient algorithms (see Gervini (2008)) for sparse functional data. Note however that these algorithms may be time-consuming, especially if both the population size and the number of measurement instants are very large. To cope with this issue, Cardot et al. (2013a) suggest in a recent work to use recursive algorithms that are very fast and allow to compute the median when the data arrive sequentially. Alternatively, Chaouch and Goga (2012) suggested to employ a weighted estimator of the L_1-median curve obtained by using only a sample drawn randomly from the population.

Considering again the electricity data presented in the Introduction we have plotted in Figure 2 the mean population curve as well as the L_1-median curve. As it can be seen, the median curve presents the same periodic patterns as the mean curve but with lower values.

3. The Horvitz-Thompson estimator and substitution estimators for functional parameters

We consider a sample s drawn from U according to a fixed-size sampling design $p(.)$, where $p(s)$ is the probability of drawing the sample s. The size n of s is nonrandom and we suppose that the first and second order inclusion probabilities satisfy $\pi_k = \mathbb{P}(k \in s) > 0$, for all $k \in U$, and $\pi_{kl} = \mathbb{P}(k \,\&\, l \in s) > 0$ for all $k, l \in U$, $k \neq l$, so that each unit and each pair of units can be drawn with a non null probability from the population.

Without any auxiliary information, Cardot et al. (2010a) proposed to estimate the total curve t_Y by the (functional) Horvitz-Thompson estimator defined as follows:

$$\hat{t}_{Y\pi} = \sum_{k \in s} \frac{Y_k}{\pi_k} = \sum_{k \in U} \frac{Y_k}{\pi_k} I_k, \tag{9}$$

where I_k is the sample membership indicator with $I_k = 1$ if $k \in s$ and $I_k = 0$ otherwise. The estimator $\hat{t}_{Y\pi}$ belongs to $L^2[0, \mathscr{T}]$ and its value at instant t, for $t \in [0, \mathscr{T}]$, is simply

$$\hat{t}_{Y\pi}(t) = \sum_{k \in s} \frac{Y_k(t)}{\pi_k}.$$

 ISSN: 2102-6238

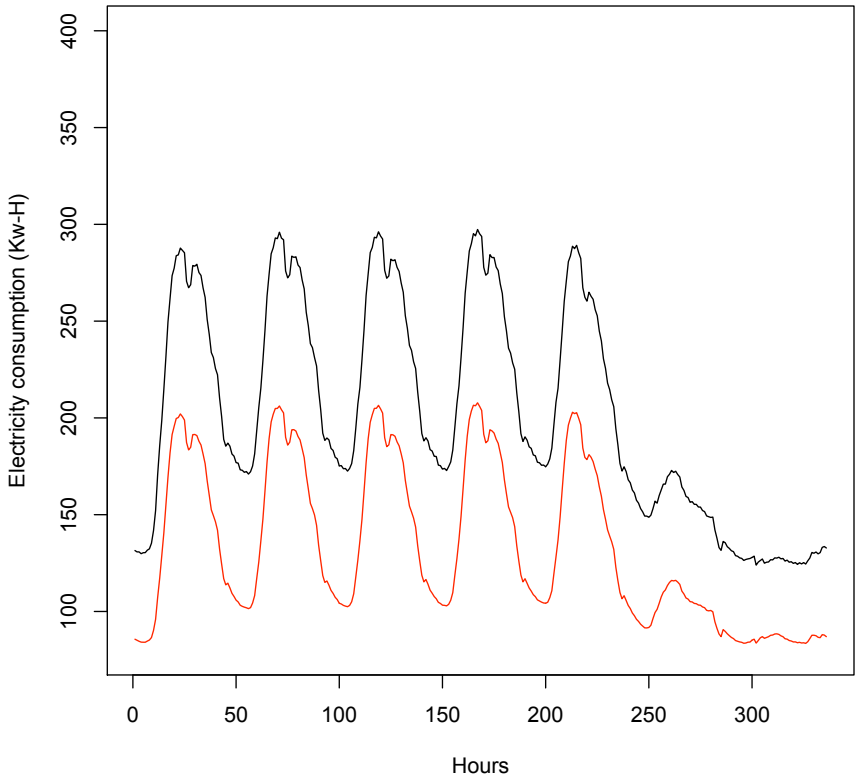


FIGURE 2. *The L_1-median profile (in red) and the mean profile (in black) of the electricity consumption curves.*

Let us remark that the curves $Y_k(t)$ are considered as fixed with respect to the sampling design and it is the sample membership I_k that is random with respect to $p(\cdot)$. Using the fact that $\mathbb{E}_p(I_k) = \pi_k > 0$, where $\mathbb{E}_p[\cdot]$ is the expectation with respect to the sampling design, we obtain easily that $\hat{t}_{Y\pi}$ is design-unbiased for t_Y, namely $\mathbb{E}_p(\hat{t}_{Y\pi}) = t_Y$. The Horvitz-Thompson estimator of the mean curve μ_N is

$$\widehat{\mu} = \frac{1}{N}\hat{t}_{Y\pi}. \tag{10}$$

Some properties of this functional estimator have been studied in Cardot et al. (2013c) and Cardot et al. (2013d). Note that the mean curve may also be estimated by the Hájek-type estimator defined as follows (see Cardot et al. (2010a), Hájek (1971)):

$$\widehat{\mu}_{Haj} = \frac{\hat{t}_{Y\pi}}{\hat{N}}, \tag{11}$$

where $\hat{N} = \sum_s 1/\pi_k$ is the Horvitz-Thompson estimator of N. This estimator may have better performances than the Horvitz-Thompson estimator $\widehat{\mu}$ in certain conditions.

 ISSN: 2102-6238

The covariance between $\hat{t}_{Y\pi}(r)$ and $\hat{t}_{Y\pi}(t)$ computed with respect to the sampling design is derived easily by using the fact that $\mathrm{Cov}_p(I_k, I_l) = \pi_{kl} - \pi_k\pi_l$ and it is given by a Horvitz-Thompson variance-type formula:

$$\gamma_p(r,t) = \sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)\frac{Y_k(r)}{\pi_k}\frac{Y_l(t)}{\pi_l}, \quad r,t \in [0,\mathscr{T}]. \tag{12}$$

For $r = t$, we obtain the variance of $\hat{t}_{Y\pi}(r)$. As $\pi_{kl} > 0$ for all $k,l \in U$, the covariance function $\gamma_p(r,t)$ is estimated unbiasedly with respect to the sampling design by:

$$\hat{\gamma}_p(r,t) = \sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{Y_k(r)}{\pi_k}\frac{Y_l(t)}{\pi_l}, \quad r,t \in [0,\mathscr{T}]. \tag{13}$$

3.1. *Taking discretization effects into account*

With real data, we generally do not observe $Y_k(t)$ at all instants t in $[0,\mathscr{T}]$ but only for a finite set of D_N measurement times, $0 = t_1 < \ldots < t_{D_N} = \mathscr{T}$. In functional data analysis, when the noise level is low and the grid of discretization points is fine, it is usual to perform a linear interpolation or to smooth the discretized trajectories in order to obtain approximations of the trajectories at every instant t (see Ramsay and Silverman (2005)). Note that the discretization points are not required to be the same for all curves. When there are no measurement errors and when the trajectories are regular enough, it is shown in Cardot and Josserand (2011), under weak regularity conditions, that linear interpolation can provide sufficiently accurate approximations of the trajectories to get efficient estimators of the mean trajectories. Thus, for each unit k in the sample s, we build the interpolated trajectory

$$Y_{k,d}(t) = Y_k(t_i) + \frac{Y_k(t_{i+1}) - Y_k(t_i)}{t_{i+1} - t_i}(t - t_i), \quad t \in [t_i, t_{i+1}], \tag{14}$$

and estimators can be constructed based on the interpolated values. For example, the Horvitz-Thompson estimator of t_Y based on the discretized observations is as follows:

$$\widehat{t}_{Y\pi,d} = \sum_{k\in s}\frac{Y_{k,d}}{\pi_k},$$

and an estimator of the mean is obtained from relation (10):

$$\hat{\mu}_d = \frac{1}{N}\hat{t}_{Y\pi,d}.$$

The covariance is then estimated by

$$\hat{\gamma}_d(r,t) = \sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{Y_{k,d}(r)}{\pi_k}\frac{Y_{l,d}(t)}{\pi_l}, \quad r,t \in [0,\mathscr{T}]. \tag{15}$$

When the observations are corrupted by noise, Cardot et al. (2013b) proposed to replace the interpolation step by a smoothing step based on local polynomials. The smoothness of the mean

 ISSN: 2102-6238

estimator depends on a bandwidth whose value is selected by a cross-validation method that accounts for the sampling weights. They have shown on simulations that smoothing does really improve the accuracy of the Horvitz-Thompson estimator only when the noise level is high. On the other hand, smoothing can lead, for low and moderate levels of noise, to estimators that are outperformed by linear interpolation methods, specially when the value of the bandwidth is not selected effectively (for instance, by usual cross-validation).

3.2. Non-linear parameters

The mean trajectory μ_N or the variance operator Γ are ratios of two finite population totals. The eigenvalues and eigenfunctions of Γ as well as the median trajectory m_N are also non-linear functions of population totals as they are defined by the implicit equations (4) and (8), respectively. To estimate these parameters, the strategy is simple and similar to the usual one for real parameters. For a unified presentation, we use the approach suggested by Deville (1999). It consists in writing the parameter of interest as a functional T of the discrete measure M defined on $L^2[0,\mathscr{T}]$ by:

$$M = \sum_{k\in U} \delta_{Y_k},$$

where δ_{Y_k} is the Dirac function taking value 1 if $Y = Y_k$ with $k \in U$ and zero otherwise. All the non-linear parameters studied here can be written as functionals of M :

$$\mu_N = \frac{\int Y dM}{\int dM}, \tag{16}$$

$$\Gamma = \frac{\int (Y-\mu_N)\otimes(Y-\mu_N) dM}{\int dM}. \tag{17}$$

The eigenvalues and eigenfunctions of Γ are also functionals of M as they are defined by the implicit equation (4). As for the median curve, we consider the functional equal to the (Fréchet) derivative with respect to y of the objective function defined in (7):

$$T_{m_N}(M;y) = -\int \frac{Y-y}{||Y-y||} dM. \tag{18}$$

Then, median is the unique solution of the implicit equation $T_{m_N}(M;m_N) = 0$. Estimators of μ, Γ and of m_N respectively, are obtained by replacing M with :

$$\hat{M} = \sum_{k\in s} \frac{\delta_{Y_k}}{\pi_k}.$$

These estimators are also called substitution estimators. More exactly, the mean trajectory μ_N is estimated by the Hájek-type estimator given in relation (11) and the variance operator Γ is estimated by

$$\widehat{\Gamma} = \frac{1}{\hat{N}} \sum_{k\in s} \frac{(Y_k - \widehat{\mu}_{Haj})\otimes(Y_k-\widehat{\mu}_{Haj})}{\pi_k} = \frac{1}{\hat{N}} \sum_{k\in s} \frac{Y_k \otimes Y_k}{\pi_k} - \widehat{\mu}_{Haj}\otimes\widehat{\mu}_{Haj}.$$

 ISSN: 2102-6238

The estimators $\hat{\lambda}_j$ for λ_j, and $\hat{v}_j$ for v_j are the eigenvalues and eigenfunctions of $\widehat{\Gamma}$, namely

$$\widehat{\Gamma}\hat{v}_j(t) = \hat{\lambda}_j \hat{v}_j(t), \quad t \in [0, \mathscr{T}]. \tag{19}$$

Considering now the median curve and assuming that all the Y_k, for $k \in s$ are not concentrated on a line, we obtain with (18), that m_N is estimated by $\hat{m}$, the unique solution of

$$\sum_{k \in s} \frac{1}{\pi_k} \frac{Y_k - \hat{m}}{\|Y_k - \hat{m}\|} = 0, \tag{20}$$

provided that we don't have $Y_k = \hat{m}$ for any $k \in s$ (see Chaouch and Goga (2012)).

4. Some asymptotic properties

We briefly present in this section some asymptotic properties of the different estimators defined before. We consider for that the asymptotic framework introduced by Isaki and Fuller (1982) and a sequence of growing and nested populations U_N of size N tending to infinity. A sample s_N of size n_N growing to infinity is drawn from U_N according to the sampling design $p_N(.)$. The first and second order inclusion probabilities are respectively denoted by π_{kN} and π_{klN}. For simplicity of notations and when there is no ambiguity, we drop the subscript N. We start by giving the asymptotic properties of an estimator $\hat{t}_{Y\pi,d}$ of the total curve t_Y or equivalently, of the Horvitz-Thompson estimator of μ_N. We present next the asymptotic properties of the non-linear estimators presented in Section 3.2. The following assumptions are needed.

Assumptions on the sampling design

A1. Assume that $\lim_{N\to\infty} \frac{n}{N} = \pi \in (0,1)$.

A2. Assume that $\min_{k\in U} \pi_k \geq \lambda > 0$, $\min_{k\neq l} \pi_{kl} \geq \lambda^* > 0$ and $\limsup_{N\to\infty} n \max_{k\neq l\in U} |\pi_{kl} - \pi_k\pi_l| < C_1 < \infty$.

A3. Assume that $\pi_{kl} = \pi_k\pi_l \left\{1 - \frac{(1-\pi_k)(1-\pi_l)}{d(\pi)}[1+o(1)]\right\}$ and $d(\pi) = \sum_{k\in U} \pi_k(1-\pi_k) \to +\infty$ as N tends to infinity.

A4. Assume that $\lim_{N\to\infty} \max_{(k,l,k',l')\in D_{4,n}} |\mathbb{E}_p\{(I_kI_l - \pi_{kl})(I_{k'}I_{l'} - \pi_{k'l'})\}| = 0$ where $D_{4,n}$ is the set of all distinct 4-tuples from U_N.

Assumptions (A1) and (A2) are classical hypotheses in survey sampling and deal with the first and second order inclusion probabilities but exclude situations in which the sampling fraction is negligible. They are satisfied for many usual sampling designs with fixed size (see for example Robinson and Särndal (1983) and Breidt and Opsomer (2000)) such as the SRSWOR and stratified designs, but the condition on π_{kl} is not satisfied for the systematic sampling design. Assumption (A2) may be weakened as seen in Breidt and Opsomer (2008) including cluster sampling design. Note however that the rates of convergence of the Horvitz-Thompson estimator are generally slower. Assumptions (A3) and (A4) are stronger and are related to variance estimation. They ensure that the sample membership indicators I_k are not too far from being independent and are satisfied for sampling designs with high entropy (see Hájek (1964), Hájek (1981), or Cardot et al. (2014)) such as SRSWOR, stratified sampling and the Poisson sampling conditioned to size or the Sampford-Durbin sampling.

 ISSN: 2102-6238

Assumptions on the regularity of trajectories

A5. There are two positive constants C_2 and C_3 and $1 \geq \beta > 1/2$ such that, for all N and for all $(r,t) \in [0,\mathscr{T}] \times [0,\mathscr{T}]$,

$$\frac{1}{N}\sum_{k\in U} Y_k(0)^2 < C_2 \quad \text{and} \quad \frac{1}{N}\sum_{k\in U}\{Y_k(t)-Y_k(r)\}^2 < C_3|t-r|^{2\beta}.$$

A6. There are two positive constants C_4 and C_5 and $1 \geq \beta > 1/2$ such that, for all N and for all $(r,t) \in [0,\mathscr{T}] \times [0,\mathscr{T}]$,

$$\frac{1}{N}\sum_{k\in U} Y_k(0)^4 < C_4 \quad \text{and} \quad \frac{1}{N}\sum_{k\in U}\{Y_k(t)-Y_k(r)\}^4 < C_5|t-r|^{4\beta}.$$

Assumptions (A5) and (A6) deal with the regularity of the trajectories, which are supposed to satisfy some moment as well as some Hölder conditions. The fact that $\beta > 1/2$ is required to get the uniform consistency (see *e.g.* the discussion in Cardot et al. (2013d)). This smoothness constraint is rather weak and do not impose that the trajectories are differentiable.

4.1. Uniform consistency of an estimator of the total or the mean curve

For each fixed value of $t \in [0,\mathscr{T}]$, the estimator $\hat{t}_{Y\pi,d}(t)$ is simply the estimator of a total of a real variable, so that under assumptions (A1) and (A2) and under the moment condition in (A5), it is consistent for t_Y (see *e.g.* Fuller (2009) and the references therein), namely

$$\text{for all} \quad \varepsilon > 0, \quad \lim_{N\to\infty} \mathbb{P}\left(\frac{1}{N}|\hat{t}_{Y\pi,d}(t) - t_Y(t)| > \varepsilon\right) = 0,$$

as well as asymptotically Gaussian,

$$\frac{\sqrt{n}}{N}(\hat{t}_{Y\pi,d}(t) - t_Y(t)) \to \mathscr{N}(0,\tilde{\gamma}_p(t))$$

where $\tilde{\gamma}_p(t) = \lim_{N\to\infty} \frac{n}{N^2}\gamma_p(t,t)$ with $\gamma_p(t,t)$ given by (12) for $r = t$. In a functional setting, it is often interesting to get the uniform consistency, namely:

$$\text{for all} \quad \varepsilon > 0, \quad \lim_{N\to\infty} \mathbb{P}\left(\sup_{t\in[0,\mathscr{T}]} \frac{1}{N}|\hat{t}_{Y\pi,d}(t) - t_Y(t)| > \varepsilon\right) = 0.$$

If assumptions (A1)-(A2) and (A5) hold and if the discretization scheme satisfies

$$\max_{i=\{1,..,D_N-1\}} |t_{i+1} - t_i|^{2\beta} = o(n^{-1}), \tag{21}$$

then, it is proven in Cardot and Josserand (2011) that the estimator $\hat{t}_{Y\pi,d}(t)$ satisfies:

$$\mathbb{E}_p\left\{\sup_{t\in[0,\mathscr{T}]} \frac{1}{N}|\hat{t}_{Y\pi,d}(t) - t_Y(t)|\right\} = O(n^{-1/2}),$$

 ISSN: 2102-6238

namely, it is asymptotically design-unbiased and uniformly consistent. Note that condition (21) ensures that the interpolation error is negligible compared to the sampling error. Under the additional assumptions (A4) and (A6) on the regularity of the trajectories and on the fourth-order inclusion probabilities, it has been shown that the variance function estimator $\hat{\gamma}_d$ given by (15) is uniformly consistent:

$$\mathbb{E}_p\left(\sup_{t\in[0,\mathscr{T}]}\frac{1}{N^2}|\hat{\gamma}_d(t,t)-\gamma_p(t,t)|\right)=o(n^{-1}).$$

4.2. *Confidence bands for the mean curve*

For each fixed measurement point $t\in[0,\mathscr{T}]$, it is possible to construct pointwise confidence intervals for $\mu_N(t)$ by using the pointwise asymptotic normality of $\hat{\mu}_d(t)$:

$$\mathbb{P}\left(\mu_N(t)\in\left[\widehat{\mu}_d(t)\pm q_\alpha\frac{\widehat{\sigma}(t)}{\sqrt{n}}\right]\right)=1-\alpha,\forall t\in[0,\mathscr{T}],$$

where $\alpha\in(0,1)$ and q_α is the quantile of order $1-\alpha/2$ of the standard normal distribution $\mathscr{N}(0,1)$. In a functional setting, we aim at building simultaneous confidence bands for μ_N of the form

$$\mathbb{P}\left(\mu_N(t)\in\left[\widehat{\mu}_d(t)\pm c_\alpha\frac{\widehat{\sigma}(t)}{\sqrt{n}}\right],\forall t\in[0,\mathscr{T}]\right)=1-\alpha, \tag{22}$$

where the coefficient c_α is unknown and depends on the desired level of confidence $1-\alpha$, and $\widehat{\sigma}(t)=\sqrt{\frac{n}{N^2}\hat{\gamma}_d(t,t)}$.

The calculation of c_α is based on the asymptotic distribution of $\hat{\mu}_d$ which has been studied in Cardot and Josserand (2011). Assuming the pointwise asymptotic normality of $\hat{\mu}_d$ and supposing that there is some $\delta>0$ such that $\lim_{N\to\infty}N^{-1}\sum_{k\in U}Y_k^{2+\delta}(t)<\infty$ for all $t\in[0,\mathscr{T}]$, it can be shown, if the discretization points are numerous enough (see condition (21)), that

$$\sqrt{n}(\hat{\mu}_d-\mu_N)\to Z\quad\text{in distribution in}\quad C[0,\mathscr{T}]$$

where Z is a Gaussian random function taking values in $C[0,\mathscr{T}]$ with mean 0 and covariance function $\tilde{\gamma}_p(r,t)=\lim_{N\to\infty}\frac{n}{N^2}\gamma_p(r,t)$. Thus, for n large enough, we have that

$$\begin{aligned}\mathbb{P}\left(\mu_N(t)\in\left[\widehat{\mu}_d(t)\pm c_\alpha\frac{\widehat{\sigma}(t)}{\sqrt{n}}\right],\forall t\in[0,\mathscr{T}]\right)&\simeq\mathbb{P}\left(\sup_{t\in[0,\mathscr{T}]}\frac{|\widehat{Z}(t)|}{\widehat{\sigma}(t)}\le c_\alpha\right)\\&\simeq\mathbb{P}\left(\sup_{t\in[0,\mathscr{T}]}\frac{|Z(t)|}{\sigma(t)}\le c_\alpha\right)\end{aligned}$$

where $\widehat{Z}$ is a zero mean Gaussian random function with covariance $\frac{n}{N^2}\hat{\gamma}_d$. The cut-off point c_α is the quantile of order $1-\alpha$ of $\sup_{t\in[0,\mathscr{T}]}|\widehat{Z}(t)|/\widehat{\sigma}(t)$ which can not be computed exactly since the distribution of the supremum of Gaussian processes is known only for few particular cases.

In a recent work, Cardot et al. (2013c) compared two methods for estimating the unknown cut-off point c_α. The first method relies on simulation of Gaussian processes and has been used

 ISSN: 2102-6238

in Degras (2011) in a non-sampling setting. This Monte Carlo method consists in simulating a Gaussian process $\widehat{Z}$ with zero mean and covariance function equal to $\hat{\gamma}_d$ in order to determine the distribution of its supremum and then estimate c_α. A rigorous mathematical justification of this technique has been given in Cardot et al. (2013b), Cardot et al. (2013d) and Cardot et al. (2014).

The second method avoids the estimation of the variance of the mean estimator by using bootstrap techniques adapted to the functional case. The variance function $\gamma_p(r,t)$ and the value c_α are estimated from the bootstrap replications. The authors used the bootstrap suggested by Gross (1980) for simple random sampling and its extensions to other sampling designs suggested by Chauvet (2007).

Using a slightly different population of load curves, Cardot et al. (2013c) compared these two methods for computing the value of c_α. They conclude that the two methods give similar coverage rates which are very close to the desired nominal rates but that the bootstrap method is much slower.

4.3. Some consistency results for the estimators of non-linear parameters

The convergence has essentially been proven in the Hilbert space $L^2[0,\mathcal{T}]$ by extending the functional approach of Deville (1999) to this space (Cardot et al. (2010a), Chaouch and Goga (2012)). For the functionals defined above (equations 16, 17 and 18), a first-order von Mises expansion (see von Mises (1947)) of the functional T may be given as follows:

$$T(\hat{M}) = T(M) + \sum_{k\in s} \frac{u_k}{\pi_k} - \sum_{k\in U} u_k + R_T, \tag{23}$$

where R_T is the reminder associated to the functional T and u_k is the linearized variable of T. Under assumptions (A1) and (A2) and if $\sup_{k\in U} ||Y_k|| < \infty$, it is proven in Cardot et al. (2010a) that the reminder term corresponding to μ_N and Γ satisfy $R_{\mu_N} = o_p(n^{-1/2})$ and $R_\Gamma = o_p(n^{-1/2})$. If all the non null eigenvalues $\lambda_j, j = 1,\ldots,N$ are distinct, then R_{λ_j} and R_{v_j} are also of order $o_p(n^{-1/2})$. For the median curve, Chaouch and Goga (2012) use the fact that the functional T given by (18) is Fréchet differentiable with respect to M and y and they obtain that $R_T = o_p(n^{-1/2})$.

The linearized variable u_k appearing in the expansion (23) is related to the first derivative of the functional T, called also the influence function, and computed at $Y = Y_k$. If $\sup_{k\in U} ||Y_k|| < \infty$, Cardot et al. (2010a) prove that the influence function of Γ exists and that the linearized variable is given by:

$$u_{k,\Gamma} = \frac{1}{N}((Y_k - \mu_N)\otimes(Y_k - \mu_N) - \Gamma), \quad k \in U.$$

If moreover, the nonnull eigenvalues of Γ are distinct, then the linearized variables of λ_j and v_j are:

$$u_{k,\lambda_j} = \frac{1}{N}(< Y_k - \mu_N, v_j >^2 - \lambda_j), \quad j = 1,\ldots N$$

$$u_{k,v_j} = \frac{1}{N}\left(\sum_{l\neq j} \frac{< Y_k - \mu_N, v_j >< Y_l - \mu_N, v_l >}{\lambda_j - \lambda_l} v_l\right), \quad j = 1,\ldots N$$

 ISSN: 2102-6238

for all $k \in U$. Finally, if $N^{-1}\sum_{k\in U}\|Y_k - m_N\|^{-1} < \infty$, then the linearized variable of the median curve is given by (see Chaouch and Goga (2012)):

$$u_{k,m_N} = \Delta^{-1}\left(\frac{Y_k - m_N}{\|Y_k - m_N\|}\right), \quad k \in U \tag{24}$$

where $\Delta = \sum_{k\in U}\frac{1}{||Y_k - m_N||}\left[\mathbf{I} - \frac{(Y_k - m_N)\otimes(Y_k - m_N)}{||Y_k - m_N||^2}\right]$ is the Jacobian operator of the functional T_{m_N} from equation (18) with $\mathbf{I}$ the identity operator defined by $\mathbf{I}y = y$.

One can remark that the linearized variables u_k are not even known for the sampled individuals, so we need to estimate them. Moreover, except in the case of the eigenvalues λ_j, u_k is a curve depending on $t \in [0, \mathscr{T}]$.

The expansion given in (23) is important since it allows approximating the substitution estimator $T(\hat{M})$ by the Horvitz-Thompson estimator of $\sum_{k\in U} u_k$, provided that the reminder term is negligible, $R_T = o_p(n^{-1/2})$. Therefore, the asymptotic variance function of the substitution estimator $T(\hat{M})$ is the Horvitz-Thompson variance:

$$AV_p(T(\hat{M}))(t) = \sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)\frac{u_k(t)}{\pi_k}\frac{u_l(t)}{\pi_l}$$

and estimated by

$$\hat{V}(T(\hat{M}))(t) = \sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{\hat{u}_k(t)}{\pi_k}\frac{\hat{u}_l(t)}{\pi_l},$$

where $\hat{u}_k(t)$ is the estimator of $u_k(t)$. In order to prove that the variance function estimator is consistent in the sense that

$$n\{\hat{V}(T(\hat{M}))(t) - AV_p(T(\hat{M}))(t)\} = o_p(1),$$

the assumption (A4) on the fourth inclusion probabilities is needed as well as additional consistency results of the linearized variable estimator $\hat{u}_k$ (Cardot et al. (2010a)). In the case of the median curve, the behavior of the variance estimator function has been studied in Chaouch and Goga (2012) by means of simulations only.

5. The particular cases of stratified and πps sampling designs

We distinguish two kinds of sampling designs, based on whether they use or do not use auxiliary information. It may happen that the auxiliary information is also a curve (for example the electricity consumption recorded during a previous period). If this information is used at the sampling stage, as in the case of stratified or proportional-to-size sampling, then the selection of the sample is more complicated than in the classical non-functional case.

Otherwise, if no auxiliary information is used at the sampling stage, the selection of the sample is realized as in the classical case. For example, a simple random sampling without replacement (SRSWOR) consists of taking n elements from the list of N elements of the population and of

 ISSN: 2102-6238

recording the curve Y_k for each sampled individual k. The Horvitz-Thompson estimator of the mean curve μ_N is $\hat{\mu} = \frac{1}{n}\sum_{k\in s} Y_k$ with the covariance function given by:

$$\gamma_{SRSWOR}(r,t) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{Y(r)Y(t),U}$$

where $S_{Y(r)Y(t),U} = \frac{1}{N-1}\sum_U (Y_k(r) - \mu_N(r))(Y_k(t) - \mu_N(t))$ is the population covariance function between $(Y_k(r))_{k\in U}$ and $(Y_k(t))_{k\in U}$. The estimator of the median is obtained from equation (20) for $\pi_k = n/N$ for all $k \in U$.

We present below two sampling designs that use auxiliary information: the stratified sampling and the proportional-to-size sampling. For each design, we give the expressions of the mean and the median curve estimators as well as their (approximated) variances. Several difficulties due to the functional nature of the data are discussed along with the suggested solutions. A small simulation study is conducted on the EDF population in order to compare these designs for estimating the mean and the median curve.

5.1. *Stratified sampling with simple random sampling within strata (STRAT)*

Suppose that the population is divided into H strata $U_1, ..., U_H$ of sizes $N_1, ..., N_H$ and sample s_h of size n_h is drawn by simple random sampling without replacement within each stratum U_h, $h = 1, ..., H$.
The mean curve estimator with stratified sampling is given by:

$$\widehat{\mu}_{\text{strat}}(t) = \sum_{h=1}^{H} \frac{N_h}{N}\left(\frac{1}{n_h}\sum_{k\in s_h} Y_k(t)\right),\ t \in [0,T], \tag{25}$$

with the covariance function given by:

$$\gamma_{\text{strat}}(r,t) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{Y(r)Y(t),U_h}\ \ r,t \in [0,\mathscr{T}], \tag{26}$$

where $S_{Y(r)Y(t),U_h} = \frac{1}{n_h-1}\sum_{k\in s_h}(Y_k(r) - \hat{\mu}_h(r))(Y_k(t) - \hat{\mu}_h(t))$ is the population covariance function between $(Y_k(r))_{k\in U}$ and $(Y_k(t))_{k\in U}$ within each stratum U_h.

Stratified sampling can be also used to estimate any non-linear parameter of interest such as the eigenvalues λ_j and the eigenfunctions v_j for $j = 1, \ldots, N$ (see Cardot et al. (2010a)), or the median curve (see Chaouch and Goga (2012)). For example, to obtain the estimator $\hat{m}_{\text{strat}}$ of the median curve with a stratified sampling, one can use the inclusion probabilities $\pi_k = n_h/N_h$ for all $k \in U_h, h = 1, \ldots H$ in equation (20) and solve the following estimation equation:

$$\sum_{h=1}^{H} \frac{N_h}{n_h}\sum_{k\in s_h} \frac{Y_k - \hat{m}_{\text{strat}}}{\|Y_k - \hat{m}_{\text{strat}}\|} = 0. \tag{27}$$

The asymptotic variance function of $\hat{m}_{\text{strat}}$ is

$$AV_{\text{strat}}(\hat{m}_{\text{strat}})(t) = \sum_{h=1}^{H} N_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right) S^2_{u_{m_N}(t),U_h}, \tag{28}$$

 ISSN: 2102-6238

where $S^2_{u_{m_N}(t),U_h}$ is the population variance function of $u_{m_N}(t) = (u_{k,m_N}(t))_{k\in U_h}$ within stratum h and u_{k,m_N} is the linearized variable of the median curve given in equation (24). That is, the lower the variation of the linearized variable within stratum, the lower the asymptotic variance of $\hat{m}_{\text{strat}}$. If the dispersion of the linearized variable within strata is indeed small, stratified sampling is efficient for estimating the median curve but may be poor for the estimation of other parameters. In such a situation, poststratification may be used (see Chaouch and Goga (2012)).

To choose the size n_h of the sample s_h, it is possible to use the proportional allocation $n_h = nN_h/N$, $h = 1,...,H$ or the optimal allocation as suggested by Cardot and Josserand (2011):

$$n_h = n\frac{N_h\sqrt{\int_0^{\mathscr{T}} S^2_{Y(r)Y(r),U_h}dr}}{\sum_{h=1}^{H} N_h\sqrt{\int_0^{\mathscr{T}} S^2_{Y(r)Y(r),U_h}dr}}, \quad h = 1,\ldots,H. \tag{29}$$

This allocation minimizes the mean variance of the stratified estimator:

$$\min_{(n_1,\ldots,n_H)} \int_0^{\mathscr{T}} \gamma_{\text{strat}}(t,t)dt \quad \text{subject to} \quad \sum_{h=1}^{H} n_h = n \quad \text{with} \quad n_h > 0, \quad \text{for } h = 1,\ldots,H.$$

This allocation is similar to that of the multivariate case when considering a total variance criterion (Cochran (1977)) and has the same interpretation, namely strata with higher variability should be sampled with a higher sampling rate than the other strata. In practice, $S^2_{Y(r)Y(r),U_h}$ are unknown for all $h = 1,...,H$. An auxiliary variable X known for all individuals $k \in U$ and highly correlated with the interest variable can be used instead and the resulting allocation is called the x-optimal allocation.

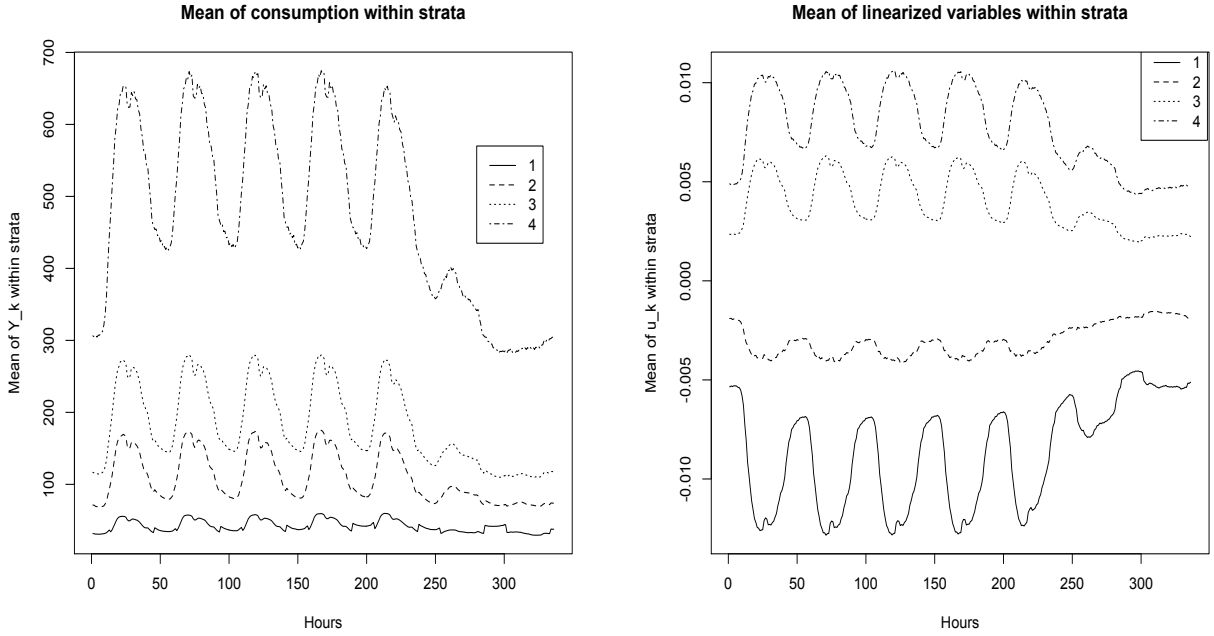
Using the allocation given by (29) may not be optimal for estimating non-linear parameters of interest. In order to derive the optimal allocation for estimating the median, for example, one should minimize the asymptotic variance of $\hat{m}_{\text{strat}}(t)$. The resulting allocation depends in this case on the linearized variable (see Chaouch and Goga (2012)).

5.2. *An illustration on load curves*

Consider the test population of $N = 18902$ French companies whose electricity consumption has been measured every half-hour over a period of two weeks. Data recorded over the first week $\mathbf{X}_k$ are used as auxiliary information, while data recorded over the second week $\mathbf{Y}_k$ are the study variable. More exactly, we have 336 instant measures per week and let $\mathbf{X}_k = (X_k(t_d))_{d=1}^{336}$ and $\mathbf{Y}_k = (Y_k(t_d))_{d=1}^{336}$. The goal is to estimate the mean curve μ_N and the median curve m_N by using a sample of size $n = 2000$ selected according to SRSWOR and STRAT designs.

The population is divided into $H = 4$ strata constructed according to the maximum level of $\mathbf{X}_k$ and based on the quartiles, so that all the strata have almost the same size (see Cardot and Josserand (2011)). The stratum 1, corresponds to consumers with low global consumption, whereas stratum 4 corresponds to consumers with high global levels of consumption. We plot in Figure 3(a), the mean of $\mathbf{Y}_k$ within each stratum and in Figure 3(b), the mean of the linearized variable of the median $\mathbf{u}_{k,m_N} = (u_{k,m_N}(t_d))_{d=1}^{336}$ within each stratum. Note that the population of the linearized variable curves is also stratified.

 ISSN: 2102-6238



(a) (b)

FIGURE 3. *Stratification based on the consumption curve: (a) Mean of the consumption curve* $\mathbf{Y}_k$ *within each stratum. (b) Mean of the linearized variable* $\mathbf{u}_{k,m_N}$ *within each stratum.*

To select a STRAT sampling, we use the proportional allocation (PROP) and the x-optimal allocation (x-OPT) computed with respect to the consumption $\mathbf{X}_k$ recorded during the previous week. Table 1 gives the size of strata and the size of samples for both types of allocation.

TABLE 1. *Strata sizes, proportional and x-optimal allocations for a sample size of* $n = 2000$.

Stratum number	1	2	3	4
Stratum size N_h	4725	4726	4725	4726
PROP allocation	500	500	500	500
x-OPT allocation	126	212	333	1329

We draw $I = 500$ samples and we give in Tables 2 and 3 statistics about the estimation errors computed according to the following loss criterion:

$$R(\hat{\theta}) = \int_0^{\mathcal{T}} |\hat{\theta}(t) - \theta(t)| dt \simeq \frac{1}{336} \sum_{d=1}^{336} |\hat{\theta}(t_d) - \theta(t_d)|,$$

with $\hat{\theta}$ an estimator of θ.

We can remark that clustering the space of functions by performing stratified sampling leads to an important gain compared to simple random sampling without replacement especially for the estimation of the mean curve. STRAT with proportional allocation gives slightly better results for

 ISSN: 2102-6238

TABLE 2. *Estimation errors of the mean curve μ_N with SRSWOR and STRAT sampling.*

	Mean	1^{st} quartile	median	3^{rd} quartile
SRSWOR	4.624	2.405	3.694	6.073
STRAT+PROP	3.731	2.116	3.041	4.803
STRAT+OPTIM	**2.507**	**1.605**	**2.198**	**3.128**

TABLE 3. *Estimation errors of the median curve m_N with SRSWOR and STRAT sampling.*

	Mean	1^{st} quartile	median	3^{rd} quartile
SRSWOR	2.697	1.362	2.274	3.527
STRAT+PROP	**1.632**	**1.048**	**1.402**	**2.017**
STRAT+OPTIM	2.263	1.444	1.969	2.865

the estimation of the median than those obtained with the optimal allocation. This is due to the fact that the optimal allocation is computed by minimizing the variance of the estimator for the mean curve; we are, however, interested in the second case in estimating the median curve.

5.3. *Probability proportional-to-size sampling: πps sampling*

Unequal probability designs are used in practice because they are usually more efficient than the equal probability designs. To estimate the mean curve, Cardot et al. (2013c) and Cardot et al. (2014) consider the fixed-size without replacement designs and to estimate the median curve, Chaouch and Goga (2012) consider with replacement probability proportional-to-size designs. We give below a description of results obtained in the first case.

For a sampling design of fixed size n, it is possible to give the equivalent of the Yates and Grundy (Yates and Grundy (1953)) and Sen formula (Sen (1953)) in the functional case. The covariance $\gamma_p(r,t)$ of $\hat{t}_{Y\pi}$ between two instants r and t, verifies

$$\gamma_p(r,t) = -\frac{1}{2}\sum_{k\in U}\sum_{l\in U, l\neq k}(\pi_{kl}-\pi_k\pi_l)\left(\frac{Y_k(r)}{\pi_k}-\frac{Y_l(r)}{\pi_l}\right)\left(\frac{Y_k(t)}{\pi_k}-\frac{Y_l(t)}{\pi_l}\right). \tag{30}$$

Using equation (30), we clearly see that the covariance $\gamma_p(r,t)$ will be small if the first-order inclusion probabilities π_k are approximately proportional to $Y_k(t)$, for all instants $t \in [0,\mathscr{T}]$. In practice, we can take π_k to be proportional to a real auxiliary variable X which is nearly proportional to the variable of interest and whose value x_k, supposed to be positive, is known for all $k \in U$. The inclusion probabilities are then given by:

$$\pi_k = n\frac{x_k}{\sum_{k\in U} x_k}. \tag{31}$$

If some x_k values are very large, it may happen that the above $\pi_k > 1$ for some elements. In this situation, we could set $\pi_k = 1$ for all k such that $nx_k > \sum_{k\in U} x_k$ and let π_k be proportional to X for the remaining elements k. Without replacement designs satisfying (31) are called *πps designs*. For given first-order inclusion probabilities π_k, there are many such sampling designs (see *e.g* Brewer and Hanif (1983) and Tillé (2006)).

 ISSN: 2102-6238

TABLE 4. *Estimation errors of the mean curve μ_N and the median curve m_N with the πps sampling.*

	Mean	1^{st} quartile	median	3^{rd} quartile
πps for μ_N	1.816	1.447	1.709	2.081
πps for m_N	7.263	2.901	5.918	9.733
πps for m_N with B-spline	**1.947**	**1.364**	**1.711**	**2.209**

Again, non-linear parameters may be estimated by using πps sampling designs. For example, the Horvitz-Thompson estimator for the median with this πps design is obtained by using in (20) the π_k given by equation (31).

5.4. An illustration on load curves

We aim at selecting a πps sample of size $n = 2000$ from the same test population of $N = 18902$ electricity curves with first-order inclusion probabilities π_k proportional to the mean consumption recorded during the previous week:

$$x_k = \frac{1}{336} \sum_{d=1}^{336} X_k(t_d), \quad k \in U. \tag{32}$$

To draw such a sample, one may use the fast version of the cube algorithm (see Chauvet and Tillé (2006)) balanced on the vector of first-order inclusion probabilities $\pi = (\pi_1, ..., \pi_N)$ with π_k given by (31) and x_k by (32). As suggested in Chauvet (2007), a random sort of the population is made before the sample selection.

We give in Table 4 the estimation errors of the mean and median curve with this πps sampling. We remark that this design performs very well for the estimation of the mean curve but very poorly for the estimation of the median. The good performance of the πps sampling for estimating the mean curve can be explained by the fact that Y_k is approximately proportional to π_k, namely

$$Y_k(t) = \pi_k \beta(t) + \varepsilon_{kt}$$

where the errors ε_{kt} are centered. In our case, the relationship between the linearized variable u_{k,m_N} and π_k is not linear as it can be remarked from (24). In order to improve the estimation of the median with a πps design, Goga suggests an estimator of m_N which consists in modifying the sampling weights $1/\pi_k$ by using a superpopulation model explaining the relationship between the u_k and π_k as follows:

$$u_{k,m_N}(t) = f(\pi_k, t) + \eta_{kt},$$

where f is unknown and the errors η_{kt} are centered. We can estimate f by using the B-spline regression as proposed by Goga and Ruiz-Gazen (2014) and obtain the following smoothed weights:

$$w_{ks} = \frac{1}{\pi_k} \left(\sum_{l \in U} \mathbf{b}'(\pi_l) \right) \left(\sum_{l \in s} \frac{\mathbf{b}(\pi_l)\mathbf{b}'(\pi_l)}{\pi_l} \right)^{-1} \mathbf{b}(\pi_k)$$

where $\mathbf{b} = (B_1, \ldots, B_q)'$ is the vector of the B-spline basis of degree m and with K interior knots, $q = K + m$. The improved estimator of the median is obtained from (20) by replacing $1/\pi_k$ with

 ISSN: 2102-6238

the weights w_{ks}. Work is actually in progress in order to obtain the asymptotic properties of this improved estimator of m_N and it will be addressed elsewhere. We give in Table 4 the estimation errors of the median estimator obtained by using the B-spline smoothed weights w_{ks} for $m = 3$, $K = 8$. We can remark that the performance of the πps design for estimating m_N was greatly improved.

5.5. *Variance estimation and confidence bands with πps sampling*

The covariance function γ_p given by (30) depends on the second-order inclusion probabilities π_{kl} which are very difficult or even impossible to calculate for many πps designs. Recently, a functional Hájek approximation for the covariance function γ_p was suggested in Cardot et al. (2013c). More exactly, suppose that the second-order inclusion probabilities satisfy the assumption (A3) given in Section 4, namely:

$$\pi_{kl} = \pi_k \pi_l \left\{ 1 - \frac{(1-\pi_k)(1-\pi_l)}{d(\pi)} [1 + o(1)] \right\}$$

where $d(\pi) = \sum_{k \in U} \pi_k (1-\pi_k)$ is supposed to tend to infinity. Then, we can approximate γ_p by the following covariance function γ_H which contains only the first-order inclusion probabilities:

$$\begin{aligned} \gamma_H(r,t) &= \sum_{k \in U} \pi_k (1-\pi_k) \left(\frac{Y_k(t)}{\pi_k} - R(t) \right) \left(\frac{Y_k(r)}{\pi_k} - R(r) \right) \\ &= \sum_{k \in U} \frac{1-\pi_k}{\pi_k} Y_k(t) Y_k(r) - \frac{1}{d(\pi)} \left(\sum_{k \in U} (1-\pi_k) Y_k(t) \right) \left(\sum_{l \in U} (1-\pi_l) Y_l(r) \right), \quad r,t \in [0, \mathscr{T}], \end{aligned} \tag{33}$$

where $R(t) = \dfrac{\sum_{k \in U} Y_k(t)(1-\pi_k)}{d(\pi)}$. This approximation appears to be very efficient when the sample size is large enough and the entropy of the sampling design is close to the maximum entropy, in particular for the rejective sampling and the Sampford-Durbin sampling (see Cardot et al. (2014)).

Using a slightly different population of load curves, the following estimator of the covariance function has been successfully used by Cardot et al. (2013c) to build confidence bands for the mean curve estimator:

$$\begin{aligned} \hat{\gamma}^*_{H,d}(r,t) &= \sum_{k \in s} (1-\pi_k) \left(\frac{Y_{k,d}(t)}{\pi_k} - \hat{R}(t) \right) \left(\frac{Y_{k,d}(r)}{\pi_k} - \hat{R}(r) \right) \\ &= \sum_{k \in s} \frac{1-\pi_k}{\pi_k^2} Y_{k,d}(t) Y_{k,d}(r) - \frac{1}{\hat{d}(\pi)} \left(\sum_{k \in s} \frac{1-\pi_k}{\pi_k} Y_{k,d}(t) \right) \left(\sum_{l \in s} \frac{1-\pi_l}{\pi_l} Y_{l,d}(r) \right), \quad r,t \in [0, \mathscr{T}], \end{aligned} \tag{34}$$

where $\hat{R}(t) = \sum_{k \in s} \dfrac{Y_{k,d}(t)(1-\pi_k)}{\pi_k} / \hat{d}(\pi)$ and $\hat{d}(\pi) = \sum_{k \in s} (1-\pi_k)$. The simulation study has shown that the confidence bands have the desired coverage rates and their widths were greatly reduced

 ISSN: 2102-6238

compared to the ones obtained with simple random sampling without replacement. The estimator (34) is the functional version of the variance estimator suggested by Deville and Tillé (2005). Consider also the following covariance estimator:

$$\hat{\gamma}_{H,d}(r,t) = \frac{\hat{d}(\pi)}{d(\pi)} \hat{\gamma}^*_{H,d}(r,t), \tag{35}$$

which is a slightly modified functional analogue of the variance estimator proposed by Berger (1998) in the real case. Assuming assumptions (A1) and (A2), it can be easily proven that $\lim_{N\to\infty} \frac{\hat{d}(\pi)}{d(\pi)} = 1$. This results implies that the covariance estimators $\hat{\gamma}^*_{H,d}$ and $\hat{\gamma}_{H,d}$ have the same asymptotic behavior.

Under assumptions (A1)-(A6) and if the discretization scheme satisfies $\lim_{N\to\infty} \max_{i\in\{1,\ldots,D_N-1\}} |t_{i+1} - t_i| = 0$, the estimator $\hat{\gamma}_{H,d}(t,t)$ is uniformly consistent for $\gamma_p(t,t)$ as shown by Cardot et al. (2014):

$$\lim_{N\to\infty} n\, \mathbb{E}_p \left(\sup_{t\in[0,\mathscr{T}]} \frac{1}{N^2} |\hat{\gamma}_{H,d}(t,t) - \gamma_p(t,t)| \right) = 0.$$

In particular, they note that the errors due to the Hájek approximation is negligible. By using the approximations of the multiple inclusion probabilities given by Boistard et al. (2012), a sharper result can be obtained for the rejective sampling:

$$\mathbb{E}_p \left(\frac{1}{N^2} (\hat{\gamma}_{H,d}(r,t) - \gamma_p(r,t)) \right)^2 = O(n^{-3}).$$

The accuracy of the proposed variance estimators has been evaluated by Cardot et al. (2014) on the population of load curves considered before. They notice that even if this estimator generally provides good estimations of the true covariance function, for a few "bad" samples, its performances could very poor. These bad performances, which fortunately occur in very rare occasions, are in fact due to a few individuals in the population that have both a very small inclusion probability π_k and a high consumption level Y_k. Further work is needed in order to build modified variance estimators that are more robust to the presence of influential individuals. More work is also needed to test the performance of this variance estimator in the case of non-linear parameters.

6. Using auxiliary information at the estimation stage

In a recent work, Cardot et al. (2013d) suggested to improve the accuracy of the Horvitz-Thompson estimator $\widehat{\mu}$ of the mean curve $\mu_N(t)$ by using a model-assisted estimator based on a functional linear model (see Faraway (1997)). This estimator can be seen as a direct extension, to the functional context, of the generalized regression estimator or GREG estimator studied in Robinson and Särndal (1983) and Särndal et al. (1992). Its main advantage is that it only requires the knowledge of the total of the auxiliary variable at the population level.

Let $X_1,\ldots,X_p$ be p real auxiliary variables and let also $\mathbf{x}_k = (X_{k1},\ldots,X_{kp})'$ be the value of the vector of auxiliary variables for the k-th individual from the population. The following superpopulation model ξ, also called functional linear model (see Faraway (1997)) is introduced:

$$\xi : \quad Y_k(t) \;=\; \mathbf{x}_k'\beta(t) + \varepsilon_{kt}, \quad t \in [0,\mathscr{T}] \tag{36}$$

 ISSN: 2102-6238

where $\beta(t) = (\beta_1(t), \ldots, \beta_p(t))'$ is the vector of functional regression coefficients, ε_{kt} are independent (across units) and centered continuous time processes, $\mathbb{E}_\xi(\varepsilon_{kt}) = 0$, with covariance function $\mathrm{Cov}_\xi(\varepsilon_{kt}, \varepsilon_{kr}) = \tilde{\Gamma}(t,r)$, for $(t,r) \in [0,\mathscr{T}] \times [0,\mathscr{T}]$.

The functional model-assisted or GREG estimator for μ_N with interpolated values $Y_{k,d}$ is given by (see Cardot et al. (2013d)):

$$\widehat{\mu}_{MA,d}(t) = \frac{1}{N}\sum_{k\in U}\widehat{Y}_{k,d}(t) - \frac{1}{N}\sum_{k\in s}\frac{(\widehat{Y}_{k,d}(t) - Y_{k,d}(t))}{\pi_k}, \tag{37}$$

where $\widehat{Y}_{k,d}(t) = \mathbf{x}_k'\widehat{\beta}_{a,d}(t)$, $t \in [t_i, t_{i+1}]$, $\widehat{\beta}_{a,d}(t) = \widehat{\mathbf{G}}_a^{-1}\frac{1}{N}\sum_{k\in s}\frac{\mathbf{x}_k Y_k(t)}{\pi_k}$ and $i = 1, \ldots, D_N$. Here, $\widehat{\mathbf{G}}_a$ is a regularized estimator of $\mathbf{G} = \frac{1}{N}\sum_{k\in U}\mathbf{x}_k\mathbf{x}_k'$ (see Cardot et al. (2013d)).

It is shown in Cardot et al. (2013d) that, under assumptions (A1)-(A2), (A5) and additional assumption on the moments of the auxiliary variables, the estimator $\widehat{\mu}_{MA,d}$ converges uniformly to the mean curve μ_N. Moreover, they also prove that

$$\widehat{\mu}_{\mathrm{MA},d} - \mu_N = \tilde{\mu} - \mu_N + o_p(n^{-1/2}),$$

where $\tilde{\mu} = \frac{1}{N}\sum_{k\in U}\mathbf{x}_k'\tilde{\beta} - \frac{1}{N}\sum_{k\in s}\frac{\mathbf{x}_k'\tilde{\beta} - Y_k}{\pi_k}$ and $\tilde{\beta}(t) = \mathbf{G}^{-1}\frac{1}{N}\sum_{k\in U}\mathbf{x}_k Y_k(t)$. This results allows to approximate the covariance function of $\widehat{\mu}_{MA,d}$ between two instants r and t by the covariance of $\tilde{\mu}$:

$$\gamma_{\mathrm{MA}}(r,t) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)\frac{Y_k(r) - \mathbf{x}_k'\tilde{\beta}(r)}{\pi_k}\frac{Y_l(t) - \mathbf{x}_l'\tilde{\beta}(t)}{\pi_l}. \tag{38}$$

A covariance estimator is given by

$$\widehat{\gamma}_{\mathrm{MA},d}(r,t) = \frac{1}{N^2}\sum_{k,l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\cdot\frac{Y_{k,d}(r) - \mathbf{x}_k'\widehat{\beta}_{a,d}(r)}{\pi_k}\cdot\frac{Y_{l,d}(t) - \mathbf{x}_l'\widehat{\beta}_{a,d}(t)}{\pi_l}, \quad r,t \in [0,\mathscr{T}]. \tag{39}$$

It is proven in Cardot et al. (2013d) that the covariance estimator $\widehat{\gamma}_{\mathrm{MA},d}$ is consistent and the variance function estimator is uniformly convergent. Thus, under additional asymptotic normality assumptions, it is also possible to build confidence bands with the Monte Carlo procedure described in Section 4.2.

Note that previous model can be extended without difficulties for auxiliary variables that vary in time, so that we have for each unit of the sample $\mathbf{x}_k(t) = (X_{k1}(t), \ldots, X_{kp}(t))'$ for $t \in [0,\mathscr{T}]$. As in Cardot et al. (2010b) nonparametric models can also be considered by first reducing the dimension of the data with principal components, as described in Section 3.2, and then consider a single index or an additive model on the principal component scores.

7. Concluding remarks

Even if some work has already been done, there are still many fields to explore, at the frontier between survey sampling and functional data analysis, in the near future.

 ISSN: 2102-6238

So far the methods of estimation combining functional data analysis and survey sampling techniques do not take into account the presence of non-response in individual curves. Trajectories with missing observations during some intervals of time may not be so rare because of transmission problems. In order to reconstruct the missing parts of the trajectories, classical methods of imputation (see Haziza (2009) for a review) can be applied, instant by instant. The disadvantage of these methods, which are essentially univariate, is that they do not take into account the history (the temporal correlation) of the individuals. Note also that a further difficulty arises from the fact that this history can also contain non-response. A second possibility would be to apply interpolation or smoothing techniques, by adapting to a survey sampling context previous works (see Staniswalis and Lee (1998)) in nonparametric estimation, on the missing part of the trajectories. This latter approach would allow the reconstruction of the individual trajectories by taking into account not only their history but also the shape of the other trajectories. Further work is needed to build an imputation method that allows to impute the trajectories by taking into account all the points of observation of the variable of interest for each individuals in our sample as well as auxiliary information. The nearest neighbor imputation technique (see Chen and Shao (2000), Shao and Wang (2008) and Beaumont and Bocci (2009)) by its nonparametric nature and its simplicity seems to be a good candidate.

When working over a long period of study, our sampling strategy which can be chosen to be well adapted at the beginning of the period, is likely to be less effective at the end. For instance, homogeneous strata at the beginning of the period may be heterogeneous after some time. Another promising direction for future investigation would be to consider samples that can change over time. A first work by Degras (2014) clearly shows that, in the case of stratified sampling, the performance of the Horvitz-Thompson estimator can be greatly improved when the sample can vary over time.

With unequal probability sampling designs, Cardot et al. (2014) noted that the Horvitz-Thompson estimator and its covariance estimator are not robust to the presence of atypical individuals. Such outlying data may not be uncommon in large samples and another interesting direction of research would be to consider correction techniques of the samplings weights of the most influential units of the sample (see *e.g.* Beaumont and Rivest (2009)) in order to get a more stable variance estimator. Some work is also needed to adapt what already exists to the functional context.

Acknowledgements

The authors thank the two anonymous referees for their careful reading and their constructive remarks.

References

Beaumont, J.-F. and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canad. J. Statist.*, 37:400–416.

Beaumont, J.-F. and Rivest, L.-P. (2009). Dealing with outliers in survey data. In Pfeffermann, D. and Rao, C., editors, *Handbook of Statistics*, volume 29A, pages 247–279. Elsevier.

Berger, Y. (1998). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *J. of Statistical Planning and Inference*, 67:209–226.

Boistard, H., Lopuhaä, H.-P., and Ruiz-Gazen, A. (2012). Approximation of rejective sampling inclusion probabilities and application to higher order correlation. *Electronic Journal of Statistics*, 6:1967–1983.

 ISSN: 2102-6238

Breidt, F.-J. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4):1023–1053.

Breidt, F.-J. and Opsomer, J. D. (2008). Endogeous post-stratification in surveys: classifying with a sample-fitted model. *The Annals of Statistics*, 36:403–427.

Brewer, K. and Hanif, M. (1983). *Sampling with unequal probabilities*. Springer-Verlag, New York.

Brown, B. (1983). Statistical use of the spatial median. *Journal of the Royal Statistical Society, B*, 45:25–30.

Cardot, H., Cénac, P., and Zitt, P.-A. (2013a). Efficient and fast estimation of the geometric median in hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19:18–43.

Cardot, H., Chaouch, M., Goga, C., and Labruère, C. (2010a). Properties of design-based functional principal components analysis. *J. of Statistical Planning and Inference*, 140:75–91.

Cardot, H., Degras, D., and Josserand, E. (2013b). Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data. *Bernoulli*, 19:2067–2097.

Cardot, H., Dessertaine, A., Goga, C., Josserand, E., and Lardin, P. (2013c). Comparison of different sample designs and construction of confidence bands to estimate the mean of functional data: An illustration on electricity consumption. *Survey Methodology*, 39:283–301.

Cardot, H., Dessertaine, A., and Josserand, E. (2010b). Semiparametric models with functional responses in a model assisted survey sampling setting. In Lechevallier, Y. and Saporta, G., editors, *Compstat 2010*, pages 411–420. Physica-Verlag, Springer.

Cardot, H., Goga, C., and Lardin, P. (2013d). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic Journal of Statistics*, 7:562–596.

Cardot, H., Goga, C., and Lardin, P. (2014). Variance estimation and asymptotic confidence bands for the mean estimator of sampled functional data with high entropy unequal probability sampling designs. *Scandinavian J. of Statistics*, 41:516–534.

Cardot, H. and Josserand, E. (2011). Horvitz-Thompson estimators for functional data: asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, 98:107–118.

Chaouch, M. and Goga, C. (2012). Using complex surveys to estimate the L_1-median of a functional variable: application to electricity load curves. *International Statistical Review*, 80(1):40–59.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.*, 91:862–872.

Chauvet, G. (2007). *Méthodes de bootstrap en population finie*. PhD thesis, Université de Rennes 2, France.

Chauvet, G. and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21:53–61.

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *J. Official Statist.*, 16:113–132.

Chiky, R. (2009). *Résumé de flux de données distribués*. PhD thesis, Sup Telecom, Paris.

Chiky, R., Dessertaine, A., and Hébrail, G. (2008). Échantillonnage sur les flux de données : état de l'art. In Guibert, P., Haziza, D., Ruiz-Gazen, A., and Tillé, Y., editors, *Méthodes de sondages*, pages 314–318, Paris. Dunod.

Cochran, W.-G. (1977). *Sampling techniques*. John Wiley & Sons, New York, third edition.

Dauxois, J. and Pousse, A. (1976). *Les analyse factorielles en calcul des probabilités et en statistique : essai d'étude synthétique*. PhD thesis, Université Paul Sabatier, Toulouse.

Degras, D. (2011). Simultaneous confidence bands for non-parametric regression with functional data. *Statistica Sinica*, 21(4):1735–1765.

Degras, D. (2014). Rotation sampling for functional data. *Statistica Sinica*, 24:1075–1095.

Deville, J.-C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Ann. Insee*, 15:3–104.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25:193–203.

Deville, J.-C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128:569–591.

Faraway, J. (1997). Regression analysis for a functional response. *Technometrics*, 39(3):254–261.

Ferraty, F. and Romain, Y., editors (2011). *The Oxford handbook of functional data analysis*. Oxford University Press, Oxford.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis. Theory and practice*. Springer Series in Statistics. Springer, New York.

Fuller, W.-A. (2009). *Sampling Statistics*. John Wiley & Sons.

Gervini, D. (2008). Robust functional estimation using the spatial median and spherical principal components. *Biometrika*, 95:587–600.

Goga, C. (2014). Improving the estimation of the functional median using survey data and B-spline modeling.

 ISSN: 2102-6238

Unpublished Technical Report.
Goga, C. and Ruiz-Gazen, A. (2014). Efficient estimation of non-linear finite population parameters by using non-parametrics. *Journal of the Royal Statistical Society, B*, 76:113–140.
Gross, S. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*.
Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35:1491–1523.
Hájek, J. (1971). Comment on a paper by D. Basu. *Foundations of statistical inference*, page 236.
Hájek, J. (1981). *Sampling from a finite population*. Statistics: Textbooks and Monographs. Marcel Dekker Inc., New York.
Haziza, D. (2009). Imputation and inference in the presence of missing data. In Pfeffermann, D. and Rao, C., editors, *Handbook of statistics*, volume 29A, pages 215–246. Elsevier.
Ilmonen, P., Oja, H., and Serfling, R. (2012). On invariant coordinate system (ics) functionals. *International Statistical Review*, 80:93–110.
Isaki, C.-T. and Fuller, W.-A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.*, 77:49–61.
Jolliffe, I. T. (2002). *Principal component analysis*. Springer Series in Statistics. Springer-Verlag, New York, second edition.
Kemperman, J. (1987). The median of a finite measure on a banach space. *In: Dodge, Y. (Ed.), Statistical Data Analysis Based on the L_1 Norm and Related Methods, North-Holland, Amesterdam*, pages 217–230.
Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46:33–50.
Lardin, P. (2012). *Estimation de synchrones de consommation électrique et prise en compte d'information auxiliaire*. PhD thesis, Université de Bourgogne.
Lohr, S. (2012). The Age of Big Data. *The New York Times*.
Ramsay, J.-O. and Silverman, B.-W. (2005). *Functional Data Analysis*. Springer Series in Statistics, New York, second edition.
Robinson, P. M. and Särndal, C. E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya : The Indian Journal of Statistics*, 45:240–248.
Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics. Springer-Verlag, New York.
Sen, A.-R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5:119–127.
Shao, J. and Wang, H. (2008). Confidence intervals based on survey data with nearest neighbor imputation. *Statist. Sinica*, 18:281–297.
Small, C. (1990). A survey of multidimensional medians. *International Statistical Review*, 58:263–277.
Staniswalis, J.-G. and Lee, J.-J. (1998). Nonparametric regression analysis of longitudinal data. *J. Amer. Statist. Assoc.*, 93:1403–1418.
Tillé, Y. (2006). *Sampling algorithms*. Springer Series in Statistics. Springer, New York.
Vardi, Y. and Zhang, C.-H. (2000). The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA*, 97(4):1423–1426.
von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics*, 18:309–348.
Weber, A. (1909). *Uber Den Standard Der Industrien, Tubingen. English translation by C.J. Freidrich (1929). Alfred Weber's theory of location of industries.* Chicago: Chicago University Press.
Weiszfeld, E. (1937). Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tôhoku Mathematical Journal*, 43:355–386.
Yates, F. and Grundy, P.-M. (1953). Selection without replacement from within strata with probability proportional to size. *J. Royal Statist. Soc., B*, 15:235–261.

 ISSN: 2102-6238