Survey Methodology, December 2013
Vol. 39, No. 2, pp. 283-301
Statistics Canada, Catalogue No. 12-001-X

Comparison of different sample designs and construction of confidence bands to estimate the mean of functional data: An illustration on electricity consumption

Hervé Cardot, Alain Dessertaine, Camelia Goga, Étienne Josserand and Pauline Lardin[1]

Abstract

When the study variables are functional and storage capacities are limited or transmission costs are high, using survey techniques to select a portion of the observations of the population is an interesting alternative to using signal compression techniques. In this context of functional data, our focus in this study is on estimating the mean electricity consumption curve over a one-week period. We compare different estimation strategies that take account of a piece of auxiliary information such as the mean consumption for the previous period. The first strategy consists in using a simple random sampling design without replacement, then incorporating the auxiliary information into the estimator by introducing a functional linear model. The second approach consists in incorporating the auxiliary information into the sampling designs by considering unequal probability designs, such as stratified and πps designs. We then address the issue of constructing confidence bands for these estimators of the mean. When effective estimators of the covariance function are available and the mean estimator satisfies a functional central limit theorem, it is possible to use a fast technique for constructing confidence bands, based on the simulation of Gaussian processes. This approach is compared with bootstrap techniques that have been adapted to take account of the functional nature of the data.

Key Words: Bonferroni; Bootstrap; Horvitz-Thompson estimator; Covariance function; Model-assisted estimator; Functional linear model; Hájek formula.

1 Introduction

With the development of automated data acquisition processes at fine time scales, it is no longer unusual to have very large databases on phenomena that change over time. For example, in the coming years in France, approximately 30 million electric meters will be replaced by smart meters. These will be able to measure the consumption of each household and business at potentially very fine time scales (by the second or minute) and send the measurements once a day to a central server. Another example is measuring the viewership of different television channels. Boxes measure in continuous time information on whether the television is on and what channel is being viewed.

The statistical unit studied is accordingly a function (of time or space), which calls for the introduction of functional analysis tools. Although this branch of statistics has existed since the 1970s (Deville 1974), Dauxois and Pousse (1976), it truly developed during the 1990s with advances in computer technology. It has applications to various fields such as climatology, economics, remote sensing, medicine and quantitative chemistry. Readers may consult the recent references Ramsay and Silverman (2005) and Ferraty and Romain (2011) for a panorama of the different techniques and examples of applications.

1. Hervé Cardot, Université de Bourgogne, Institut de Mathématiques de Bourgogne, 9 av. Alain Savary, 21078 DIJON, FRANCE; Alain Dessertaine, LA POSTE - DIRECTION DU COURRIER - DFI – DCPES, 2 Boulevard Newton 77543 MARNE LA VALLEE CEDEX 2 and EDF, R&D, ICAME-SOAD, 1 av. du Général de Gaulle, 92141 CLAMART, France; Camelia Goga, Université de Bourgogne, Institut de Mathématiques de Bourgogne; Étienne Josserand, Université de Bourgogne, Institut de Mathématiques de Bourgogne. E-mail: camelia.goga@u-bourgogne.fr; Pauline Lardin, Université de Bourgogne, Institut de Mathématiques de Bourgogne and EDF, R&D, ICAME-SOAD.

When the potential databases are very large, it can be difficult and costly to collect, save and analyze the entire data set. Moreover, if one is interested in simple indicators such as the mean curve under constraints of memory space or the cost of transmission, the use of survey techniques to extract a sample can provide a precise estimate at a reasonable cost (Dessertaine 2008).

In the statistical literature, there are as yet few studies that combine functional data analysis and sampling theory. Cardot, Chaouch, Goga and Labruère (2010) are interested in using principal component analysis to reduce the dimension of the data, while Cardot and Josserand (2011) examine the uniform convergence properties of Horvitz-Thompson estimators of mean curves. Chaouch and Goga (2012) provide a robust estimator of central curves.

The objective of this study is to compare different sampling strategies in a functional context, using a real example. These real data concern the electricity consumption, measured every half hour for two weeks, of a test population of $N = 15,069$ electric meters. The time profile of individuals' electricity use depends on covariables such as weather conditions (temperature, *etc.*) or geographic characteristics (altitude, latitude or longitude). Unfortunately, those variables are not available for this study, and we use only one variable as auxiliary information: the mean consumption from each meter during the previous week. This information can easily be transmitted by all the meters in the population.

Extending estimation methods that use auxiliary information to the functional framework is not always straightforward. Cardot and Josserand (2011) propose stratifying the population of curves to improve the estimate of the mean curve. Chaouch and Goga (2012), who are interested in the median curve, suggest using PPS (probability proportional to size) sampling with replacement as well as a post-stratified estimator. In this article, we propose to compare several strategies that take auxiliary information into account. The first strategy uses auxiliary information in selection of the sample: sampling with an unequal probabilities design (stratified, πps) and estimation with the Horvitz-Thompson estimator. The second strategy introduces this information at the estimation stage: a simple random sample is drawn without replacement and estimation is performed using a linear regression model (Särndal, Swensson and Wretman 1992) adapted to the functional framework (Faraway 1997).

A new question, related to the functional nature of the data, naturally arises: how to quantify sampling uncertainty? The construction of confidence intervals—a central concern for survey methodologists—has received little attention in the field of functional data statistics, where it is a matter of constructing confidence bands. Drawing on techniques based on estimation of the covariance function of the estimator (see Faraway (1997), Cuevas, Febrero and Fraiman (2006) or more recently Degras (2011)), we first propose to construct confidence bands by simulating Gaussian processes. An asymptotic justification of the validity of these techniques is given in Cardot, Degras and Josserand (2013) when the hypotheses of the central limit theorem are verified and there is a precise estimator of the covariance function. A second method of construction, which is based on bootstrap techniques, is also applied. It basically consists of three bootstrap techniques for use in a finite population: the bootstrap without replacement proposed by Gross (1980), the rescaling bootstrap (Rao and Wu 1988) and the mirror-match bootstrap (Sitter 1992). In this study, we use the bootstrap without replacement, which is based on adaptations for the stratified and PPS designs proposed by Chauvet (2007).

In Section 2, we introduce notations, estimators of the mean curve where there is auxiliary information, and estimators of their covariance function. The algorithms for constructing confidence bands, based on the bootstrap or simulation of Gaussian processes, are described in Section 3. Section 4 then compares the

different strategies—in terms of precision of the estimators, width and coverage of the confidence bands and computation time—for purposes of estimating the consumption curves of the French electricity company EDF (Électricité de France). For this we use samples of size $n = 1,500$ in our test population consisting of $N = 15,069$ curves. To finish, we present several perspectives of research in Section 5.

2 Functional data in a finite population

Consider a finite population $U = \{1, ..., N\}$ of size N and assume that for each unit k in the population U, we can observe the deterministic curve $Y_k = \left(Y_k(t)\right)_{t \in [0,T]}$. The objective is to estimate the mean curve of the population, which is defined for any instant $t \in [0, T]$, by

$$\mu(t) = \frac{1}{N} \sum_{k \in U} Y_k(t).$$

Let s be a sample of fixed size n, selected randomly in U, according to a sampling design $p(.)$. Let $\pi_k = \Pr(k \in s)$ and $\pi_{kl} = \Pr(k \& l \in s)$ be the first- and second- order inclusion probabilities respectively. Assume that $\pi_k > 0$ for any unit k in population U.

The mean curve μ is estimated using the Horvitz-Thompson estimator (Cardot *et al.* 2010) as follows:

$$\hat{\mu}(t) = \frac{1}{N} \sum_{k \in s} \frac{Y_k(t)}{\pi_k} = \frac{1}{N} \sum_{k \in U} \frac{Y_k(t)}{\pi_k} 1_{k \in s}, \quad t \in [0, T], \tag{2.1}$$

where $1_{k \in s}$ is the indicator that unit k belongs to the sample s. For each instant $t \in [0, T]$, the estimator $\hat{\mu}(t)$ is unbiased for $\mu(t)$, meaning that $E\left(\hat{\mu}(t)\right) = \mu(t)$ where the expectation is considered in relation to the sampling design.

Generally, the trajectories $Y_k(t)$ are not observed continuously for $t \in [0, T]$ but only for a set of D measurement instants $0 = t_1 < t_2 < \ldots < t_D = T$. In functional data analysis, a classical strategy is to interpolate or smooth discretized trajectories to obtain objects that are truly functions (Ramsay and Silverman 2005). This also makes it possible to deal with curves whose measurement instants are not identical. In the context of surveys, Cardot and Josserand (2011) studied linear interpolation where there is no measurement error at the discretized points, while Cardot *et al.* (2013) examined smoothing procedures. If there are enough discretization points and the trajectories are fairly regular (but not necessarily derivable), the approximation error due to smoothing or interpolation is negligible in relation to the sampling error. We subsequently assume that the trajectories are observed at any point t of the interval $[0, T]$.

The Horvitz-Thompson covariance function $\gamma(r, t) = \operatorname{cov}\left(\hat{\mu}(r), \hat{\mu}(t)\right)$ is given by

$$\gamma(r, t) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{Y_k(r)}{\pi_k} \frac{Y_l(t)}{\pi_l}$$

for any $(r,t) \in [0,T] \times [0,T]$ and $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. If we assume that the second-order probabilities of inclusion satisfy $\pi_{kl} > 0$, an unbiased estimator of $\gamma(r,t)$ is given by the Horvitz-Thompson unbiased estimator of the variance,

$$\hat{\gamma}(r,t) = \frac{1}{N^2} \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{Y_k(r)}{\pi_k} \frac{Y_l(t)}{\pi_l} \tag{2.2}$$

for any $(r,t) \in [0,T] \times [0,T]$.

2.1 Using auxiliary information for estimating the mean trajectory

It is well known that using auxiliary information that effectively explains the variable of interest can greatly improve the precision of the Horvitz-Thompson estimator. In the case of the EDF data, the outside temperature or the type of contract could probably be useful auxiliary variables. A stratification based on geographic position would also yield estimates for different regions. In this study, we have as an auxiliary variable the total electricity consumption for the previous week. We assume that this variable (a real one) is observed for all units in the population.

In this section, we present the Horvitz-Thompson estimator for the mean curve as well as an estimate of the covariance function of this estimator, both for a stratified design using simple random sampling without replacement (SRSWOR) in each stratum, denoted hereafter as STRAT, and for PPS sampling without replacement, which will be denoted as πps. We also consider an estimator of the mean curve, assisted by a functional linear model.

2.1.1 Stratified sampling with SRSWOR in each stratum (STRAT)

The population U is assumed to be stratified into a fixed number H of strata $U_1, \ldots, U_H$ of sizes $N_1, \ldots, N_H$. Within each stratum U_h, a sample s_h of size n_h is drawn according to an SRSWOR design.

We denote $\mu_h(t) = \sum_{k \in U_h} Y_k(t) / N_h$, for $t \in [0,T]$, the mean curve in each stratum, and $\hat{\mu}_h(t) = \sum_{k \in s_h} Y_k(t) / n_h$, its estimate. The estimator of the mean curve μ is then defined by

$$\hat{\mu}_{\text{strat}}(t) = \frac{1}{N} \sum_{h=1}^{H} N_h \hat{\mu}_h(t) = \sum_{h=1}^{H} \frac{N_h}{N} \left(\frac{1}{n_h} \sum_{k \in s_h} Y_k(t) \right), \; t \in [0,T]. \tag{2.3}$$

The Horvitz-Thompson estimator of the covariance function γ is then

$$\hat{\gamma}_{strat}(r,t) = \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{Y(r)Y(t),s_h} \quad r,t \in [0,T], \tag{2.4}$$

where

$$S_{Y(r)Y(t),s_h} = \frac{1}{n_h - 1} \sum_{k \in s_h} \left(Y_k(r) - \hat{\mu}_h(r) \right) \left(Y_k(t) - \hat{\mu}_h(t) \right)$$

is the estimator of the covariance function $S_{Y(r)Y(t),U_h}$ in stratum h. For $r = t \in [0,T]$, we obtain the estimator of the variance function as follows:

$$\hat{\gamma}_{strat}(r) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S^2_{Y(r),s_h},$$

where

$$S^2_{Y(r),s_h} = \frac{1}{n_h - 1}\sum_{k\in s_h}\left(Y_k(r) - \hat{\mu}_h(r)\right)^2$$

is the estimator of the variance $S^2_{Y(r),U_h}$ in stratum h. Cardot and Josserand (2011) propose an extension, in the functional framework, of Neyman's optimal allocation. When the sizes n_h of the samples s_h verify

$$n_h = n\frac{N_h\sqrt{\int_0^T S^2_{Y(r),U_h}dr}}{\sum_{h=1}^{H} N_h\sqrt{\int_0^T S^2_{Y(r),U_h}dr}}, \quad h = 1,...,H, \tag{2.5}$$

the integrated variance, $\int_0^T \hat{\gamma}_{\text{strat}}(t)\,dt$, of the stratified estimator is minimized. This allocation is similar to the one obtained in a multivariate context by Cochran (1977). By replacing the variable Y by another variable X that is known for the entire population and is highly correlated with the variable of interest, we obtain an allocation that can be described as x- optimal.

Note 2.1 *For $H = 1$, we obtain the simple random design without replacement (SRSWOR), and the mean curve $\mu(t)$ is estimated by*

$$\hat{\mu}_{\text{srswor}}(t) = \frac{1}{n}\sum_{k\in s} Y_k(t), \quad t \in [0,T]. \tag{2.6}$$

The estimator of the covariance function defined in (2.2) is then

$$\hat{\gamma}_{\text{srswor}}(r,t) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{Y(r)Y(t),s}. \tag{2.7}$$

2.1.2 PPS sampling without replacement (πps)

PPS sampling designs with or without replacement are often used in practice because they are more effective than equal probability designs when the variable of interest is basically proportional to an auxiliary variable X that has strictly positive values.

In the case of samples of fixed size n drawn without replacement, it is possible to give the equivalent of the formula of Yates and Grundy (1953) and Sen (1953). The covariance function of $\hat{\mu}$ verifies

$$\gamma(r,t) = -\frac{1}{2}\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U, l\neq k}(\pi_{kl} - \pi_k\pi_l)\left(\frac{Y_k(r)}{\pi_k} - \frac{Y_l(r)}{\pi_l}\right)\left(\frac{Y_k(t)}{\pi_k} - \frac{Y_l(t)}{\pi_l}\right),\ r,t \in [0,T]. \quad (2.8)$$

Assume that the values x_k of variable X are known for all units k in the population. It is then possible to define the inclusion probabilities as follows:

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

Methods have been proposed in the literature for the case $\pi_k > 1$ (Särndal *et al.* 1992).

Second-order inclusion probabilities are generally very difficult to calculate for πps designs, and therefore Formula (2.2) cannot be used. However, there is a simple asymptotic approximation of the variance, which was proposed by Hájek (1964) and which entails only first-order inclusion probabilities. This approximation proves to be very effective when the sample is large and the entropy of the sampling design is close to maximum entropy. To select sample s with inclusion probabilities π_k, the cube algorithm (Deville and Tillé 2004) balanced on the variable $\boldsymbol{\pi} = (\pi_k)_{k\in U}$ can be used. Deville and Tillé (2005) show that for this particular sampling design, the Hàjek formula is highly effective for estimating the variance of a total or a mean. This formula for approximating the variance can also be used for the covariance, which is then estimated by

$$\hat{\gamma}_{\pi\text{ps}}(r,t) = \frac{1}{N^2}\sum_{k\in s}(1-\pi_k)\left(\frac{Y_k(r)}{\pi_k} - \hat{R}(r)\right)\left(\frac{Y_k(t)}{\pi_k} - \hat{R}(t)\right),\ \ r,t \in [0,T], \quad (2.9)$$

where

$$\hat{R}(t) = \frac{\sum_{k\in s}\frac{Y_k(t)}{\pi_k}(1-\pi_k)}{\sum_{k\in s}(1-\pi_k)}.$$

We also used the systematic sampling with unequal probabilities proposed by Madow (1949), since it is simple to use. Unfortunately, it is difficult to estimate the variance for this type of design, and we will therefore not use it to construct confidence bands.

2.2 The model-assisted estimator

Consider p real auxiliary variables $X_1,\ldots,X_p$ and let x_{kj} be the value of the variable X_j for the k^{th} individual. Let $\mathbf{x}_k = (x_{k1},\ldots,x_{kp})'$ denote the vector containing the values of p auxiliary variables measured on the k^{th} individual. We consider that the relationship between the variable of interest and the auxiliary variables is modeled by the following superpopulation model

$$\xi : Y_k(t) = \mathbf{x}_k'\boldsymbol{\beta}(t) + \varepsilon_{kt},\ \ t \in [0,T] \quad (2.10)$$

with

$$E_\xi\left(\varepsilon_{kt}\right) = 0, E_\xi\left(\varepsilon_{kt}\varepsilon_{lt'}\right) = 0 \text{ for } k \neq l \text{ and } E_\xi\left(\varepsilon_{kt}\varepsilon_{kt'}\right) = \sigma_{tt'}^2 \text{ for } k = l.$$

This model is an immediate generalization of the functional linear model proposed by Faraway (1997) to several auxiliary variables.

The estimate of $\boldsymbol{\beta}$ based on the model ξ and the sampling design $p(.)$ is given by

$$\hat{\boldsymbol{\beta}}(t) = \left(\sum_{k\in s}\frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k}\right)^{-1}\sum_{k\in s}\frac{\mathbf{x}_k Y_k(t)}{\pi_k}, \quad t \in [0,T]. \tag{2.11}$$

Note that the sampling weights do not depend on the time $t \in [0,T]$. Let $\hat{Y}_k(t) = \mathbf{x}_k'\hat{\boldsymbol{\beta}}(t)$ be the estimator based on the sampling design for the prediction of $Y_k(t)$ under the model ξ. By direct analogy with the univariate case (Särndal *et al.* 1992), we finally obtain the following estimator for the mean, for $t \in [0,T]$,

$$\begin{aligned}\hat{\mu}_{MA}(t) &= \frac{1}{N}\sum_{k\in s}\hat{Y}_k(t) - \frac{1}{N}\sum_{k\in s}\frac{\left(\hat{Y}_k(t) - Y_k(t)\right)}{\pi_k} \\ &= \frac{1}{N}\sum_{k\in U}\frac{Y_k(t) - \mathbf{x}_k'\hat{\boldsymbol{\beta}}(t)}{\pi_k} + \frac{1}{N}\left(\sum_{k\in U}\mathbf{x}_k'\right)\hat{\boldsymbol{\beta}}(t).\end{aligned} \tag{2.12}$$

If the ξ contains the constant variable 1, then the estimator becomes

$$\hat{\mu}_{MA}(t) = \frac{1}{N}\sum_{k\in U}\hat{Y}_k(t), \quad t \in [0,T]. \tag{2.13}$$

For fixed r and t, the asymptotic covariance of $\hat{\mu}_{MA}(r)$ and $\hat{\mu}_{MA}(t)$ can be calculated according to the classical residual technique (Särndal *et al.* 1992),

$$\gamma_{MA}(r,t) \simeq \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\left(\pi_{kl} - \pi_k\pi_l\right)\frac{\left(Y_k(r) - \tilde{Y}_k(r)\right)}{\pi_k}\frac{\left(Y_l(t) - \tilde{Y}_l(t)\right)}{\pi_l}, \tag{2.14}$$

where $\tilde{Y}_k(r) = \mathbf{x}_k'\tilde{\boldsymbol{\beta}}(t)$ is the prediction of $Y_k(t)$ under the superpopulation model and $\tilde{\boldsymbol{\beta}}(t) = \left(\sum_U \mathbf{x}_k\mathbf{x}_k'\right)^{-1}\left(\sum_U \mathbf{x}_k Y_k(t)\right)$ is the estimate of $\boldsymbol{\beta}$ at the level of the population and $r,t \in [0,T]$. Cardot, Goga and Lardin (2013) show that this result remains valid uniformly in $r,t \in [0,T]$.

As an estimator of the covariance function $\gamma_{MA}(r,t)$, we propose the Horvitz-Thompson estimator of asymptotic covariance given by (2.14) where $\tilde{\boldsymbol{\beta}}(t)$ is replaced by its estimator $\hat{\boldsymbol{\beta}}(t)$ based on the sampling design,

$$\hat{\gamma}_{MA}(r,t) = \frac{1}{N^2}\sum_{k,l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{\left(Y_k(r) - \hat{Y}_k(r)\right)}{\pi_k}\frac{\left(Y_l(t) - \hat{Y}_l(t)\right)}{\pi_l}, \quad r,t \in [0,T]. \tag{2.15}$$

Note 2.2 *It is entirely possible to consider a superpopulation model ξ that is more general than the linear model proposed here. Estimation techniques based on smoothing by B-splines (Goga and Ruiz-Gazen 2012) can then also be considered. In our study, the relationship between consumption at instant t and the mean consumption for the previous week is almost linear (cf. Figure 4.1), which justifies not using these non-parametric approaches.*

3 Construction of confidence bands

Here we are considering confidence bands for the mean curve μ that have the form

$$\mathbb{P}\left(\mu(t) \in \left[\hat{\mu}(t) \pm c_\alpha \hat{\sigma}(t)\right],\ \forall t \in [0,T]\right) = 1-\alpha, \tag{3.1}$$

where the value of the coefficient c_α is unknown and depends on the desired confidence level $1-\alpha$, and $\hat{\sigma}(t)$ is an estimator of the standard deviation of $\hat{\mu}(t)$. The calculation of c_α is based on the fact that according to some hypotheses (Cardot *et al.* 2013), the process

$$Z(t) = (\hat{\mu}(t) - \mu(t)) / \hat{\sigma}(t),\quad t \in [0,T],$$

converges toward a Gaussian process in the space of continuous functions $\mathcal{C}([0,T])$. We then have

$$\mathbb{P}\left(\sup_{t\in[0,T]} |Z(t)| \le c_\alpha\right) = \mathbb{P}\left(\mu(t) \in \left[\hat{\mu}(t) \pm c_\alpha \hat{\sigma}(t)\right], \forall t \in [0,T]\right) \tag{3.2}$$

and it is therefore sufficient to determine c_α, the quantile of order $1-\alpha$ of the real random variable $\sup_{t\in[0,T]} |Z(t)|$ to construct the confidence band completely. The distribution of the sup of Gaussian processes is known explicitly for only a few specific cases, such as the Brownian motion.

We propose two approaches to determine the value of c_α. The first is based on a direct estimate of the standard deviation and the simulation of Gaussian processes $Z(t)$. The second, which does not require having an estimator of the variance, is based on resampling techniques where both the standard deviation and the value of c_α are obtained from bootstrap replications.

3.1 Construction of confidence bands by simulation of Gaussian processes

The steps of the algorithm are as follows:

1) Draw sample s of size n using sampling design p and calculate the estimator $\hat{\mu}$ as well as the estimator $\hat{\gamma}(r,t)$ of the covariance function $\gamma(r,t)$, $r,t \in [0,T]$.
2) Simulate M curves Z_m, $m = 1,\ldots M$, of the same distribution as Z where Z is a Gaussian process of expectation 0 and of covariance function ρ where $\rho(r,t) = \hat{\gamma}(r,t) / \left(\hat{\gamma}(r)\hat{\gamma}(t)\right)^{1/2}$ $r,t \in [0,T]$.

3) Determine c_α, the quantile of order $1-\alpha$ of the variables, $\left(\sup_{t\in[0,T]}\left|Z_m(t)\right|\right)_{m=1,\ldots,M}$.

This algorithm, which is very fast and easy to implement, has already been proposed in the context of i.i.d. observations by Faraway (1997), Cuevas *et al.* (2006) and Degras (2011) to construct confidence bands. A rigorous asymptotic justification of this approach may be found in Cardot *et al.* (2013) for sampling in finite populations.

3.2 Construction of confidence bands by bootstrapping

In this work, we use the bootstrap method proposed by Gross (1980) for SRSWOR sampling and the extensions proposed by Chauvet (2007) for STRAT and πps designs. It is based on the following principle: the sample s is used to simulate a fictitious population U^* in which we select a number of bootstrapped samples. The implementation of this algorithm is not straightforward when the ratio $1/\pi_k$ is not an integer. Many variants have been proposed in the literature to deal with the general case, and we decided to adopt the one first proposed by Booth, Butler and Hall (1994) for the SRSWOR design.
Assume that sample s of size n was selected using sampling design p and let $\hat{\mu}$ be the estimator of μ calculated from s.

General bootstrap algorithm

1) Duplicate each individual $k \in s$, $\left[1/\pi_k\right]$ times, where [.] designates the integer portion. We complete the population thus obtained by selecting a sample in s with an inclusion probability $\alpha_k = 1/\pi_k - \left[1/\pi_k\right]$. Let Y_k^*, $k \in U^*$ be the value of the variable of interest in the pseudo-population.
2) Draw M samples s_m^*, $m = 1,\ldots,M$ of size n in the pseudo-population U^* using the sampling design p^* with inclusion probabilities π_k^* and calculate

$$\hat{\mu}_m^*(t) = \frac{1}{N}\sum_{k\in s_m^*}\frac{Y_k^*(t)}{\pi_k^*},\ t\in\left[0,T\right] \text{ and } m = 1,\ldots,M.$$

3) Estimate the function $\hat{\sigma}(t)$ by the corrected empirical standard deviation of $\hat{\mu}_m^*(t)$, $m = 1,\ldots,M$,

$$\hat{\sigma}^2(t) = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\mu}_m^*(t) - \hat{\mu}_\bullet^*(t)\right)^2,$$

where

$$\hat{\mu}_\bullet^*(t) = \frac{1}{M}\sum_{m=1}^{M}\hat{\mu}_m^*(t) \text{ and } t\in\left[0,T\right].$$

4) Choose c_α as the quantile of order $1-\alpha$ of the variables

$$\left(\sup_{t\in[0,T]} \frac{\mid \hat{\mu}_m^*(t) - \hat{\mu}(t) \mid}{\hat{\sigma}(t)} \right)_{m=1,\ldots,M}.$$

A technique similar to the one used in step 4 of the algorithm was used by Bickel and Krieger (1989) to construct confidence bands for a distribution function.

The SRSWOR design uses the general bootstrap algorithm for $\pi_k^* = n / N$, and for the STRAT design, we apply in each stratum U_h, for $h = 1, \ldots, H$, the algorithm used for the SRSWOR design with $\pi_k^* = n_h / N_h \;\; k \in U_h$. In this case, we are back to the algorithm proposed by Booth *et al.* (1994).

The adaptation of the bootstrap algorithm to the πps design was proposed by Chauvet (2007). It consists in selecting, during step 2 of the general algorithm, the sample s^* in U^* with inclusion probabilities

$$\pi_k^* = \frac{nx_k}{\sum_{k\in U^*} x_k}.$$

This change is necessary in order to comply with the constraint of fixed size during re-sampling. The inclusion probabilities π_k^* are also used to estimate $\hat{\mu}_m^*$ in step 2 of the general algorithm. The selection of a πps sample can be carried out using the cube algorithm with the balancing variable π. In these conditions, it is desirable to perform a random sort in the population U (or U^*) before the selection of s (or s_m^*) in order to obtain a sampling design close to maximum entropy (Chauvet 2007, Tillé 2011). Chauvet (2007) also gives asymptotic results concerning the convergence of the variance estimator obtained in the case of the bootstrap for the πps design.

Finally, it is also possible to adapt this general algorithm to estimate the variance function of the estimator $\hat{\mu}_{MA}$. In step 1 of the algorithm, we also calculate the values $\mathbf{x}_k^*$ of $\mathbf{x}_k$ in the pseudo-population U^*. Using the fact that the linear-model-assisted estimator is a nonlinear function of Horvitz-Thompson estimators, we calculate the bootstrapped value $\hat{\mu}_{MA}^*$ of $\hat{\mu}_{MA}$ over sample s_m^* according to

$$\hat{\mu}_{MA}^*(t) = \frac{1}{N} \sum_{k\in s_m^*} \frac{Y_k^*(t) - \mathbf{x}_k^{*\prime}\hat{\boldsymbol{\beta}}^*(t)}{\pi_k^*} + \frac{1}{N}\left(\sum_{k\in U} \mathbf{x}_k^{\prime} \right) \hat{\boldsymbol{\beta}}^*(t)$$

where $\hat{\boldsymbol{\beta}}^*(t) = \left(\sum_{s_m^*} \mathbf{x}_k^* \mathbf{x}_k^{*\prime} \right)^{-1} \sum_{s_m^*} \mathbf{x}_k^* Y_k^*(t)$. As Canty and Davison (1999) note, using the total of the variable $\mathbf{x}_k$ over the population U instead of the pseudo-population U^* yields better results, especially when this variable has extreme values.

4 Study of the mean electricity consumption curve

We have a population U consisting of $N = 15,069$ electricity consumption curves measured every half hour during two consecutive weeks. We have $D = 336$ measurement points for each week, and we

want to estimate the mean consumption curve for the second week. We denote by $\mathbf{Y}'_k = \left(Y_k\left(t_1\right),\ldots,Y_k\left(t_D\right)\right)$, the electricity consumption of individual $k \in U$ measured in the second week and $\mathbf{X}'_k = \left(X_k\left(t_1\right),\ldots,X_k\left(t_D\right)\right)$, the individual's consumption during the first week. The mean consumption of each individual k during the first week, $x_k = \sum_{d=1}^{D} X_k\left(t_d\right) / D$, which is simple piece of information that is inexpensive to transmit, will be used as auxiliary information. This variable (a real one), which is known for all units k in the population, is strongly related to the current consumption curve. As Figure 4.1 shows, the current consumption in each t is almost proportional to the mean consumption for the previous week.

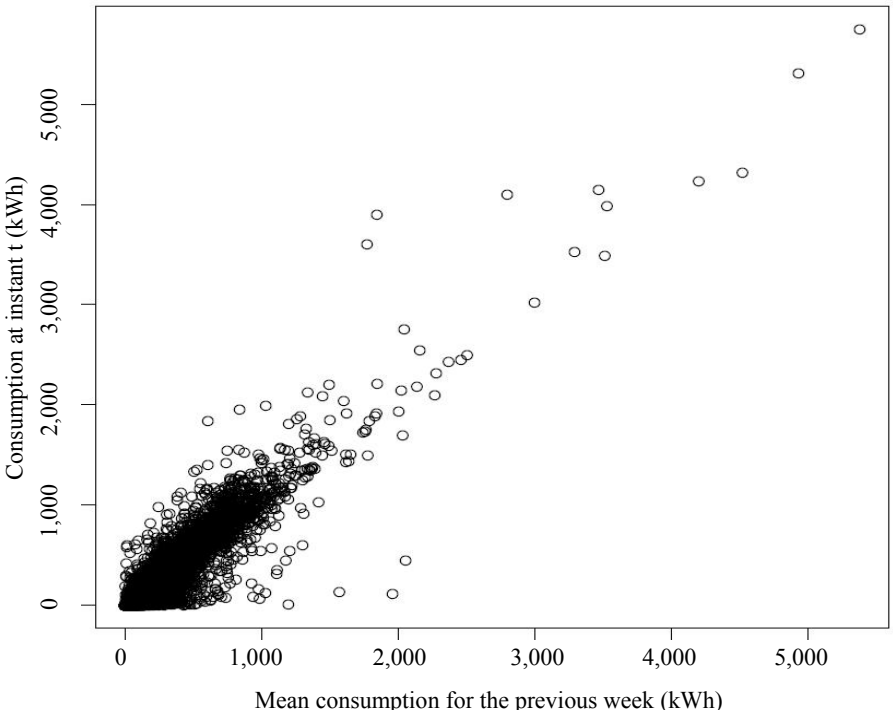


Figure 4.1 Representation of consumption at an instant t as a function of the mean consumption for the previous week

4.1 Description of strategies used

We consider samples of fixed size $n = 1,500$ obtained using different sampling designs. The strategies presented are repeated I times to evaluate and compare their performance.

1. *SRSWOR sampling and Horvitz-Thompson estimator*

This design is simple to implement; the Horvitz-Thompson estimator of the mean curve is given by (2.6) and the estimator of its covariance by (2.7).

2. *STRAT stratified design and Horvitz-Thompson estimator*

A stratified design is very effective if the strata are homogenous in relation to the variable of interest. In this study, we used the k-means algorithm to create the strata, and we considered $H = 10$ strata. A first stratification (STRAT 1) was carried out using the classification of the discretized trajectories $\mathbf{X}'_k$ for the

previous week. A second stratification, which uses only the aggregate information x_k was also considered. It is denoted by STRAT 2.

Tables 4.1 and 4.2 show the sizes of the strata N_h yielded using the two stratifications and the optimal sizes n_h, according to (2.5), of the samples to be selected in each stratum. In both cases, the strata are numbered in ascending order in relation to the mean consumption for each stratum. More specifically, stratum 1 corresponds to small consumers of electricity and stratum 10 refers to the 10 largest consumers. Note that the first stratification, which requires knowing the electricity consumption at each measurement instant t, requires more information than the second stratification. The mean curve is constructed using (2.3), and its covariance is estimated by (2.4).

Table 4.1
STRAT 1: Stratification based on curves. The strata are constructed using the curves for week 1. The optimal allocation n_h is calculated using the curves for week 1.

h	1	2	3	4	5	6	7	8	9	10
N_h	3,866	4,769	623	2,690	664	1,251	806	328	62	10
n_h	212	345	87	242	117	179	172	101	35	10

Table 4.2
STRAT 2: Stratification based on the mean consumption x_k. The optimal allocation n_h is calculated using the mean consumption for week 1.

h	1	2	3	4	5	6	7	8	9	10
N_h	3,257	4,236	3,139	1,937	1,189	731	415	125	30	10
n_h	260	293	248	204	159	133	111	56	26	10

3. *πps sampling and Horvitz-Thompson estimator*

We used the cube algorithm proposed by Deville and Tillé (2004) and Chauvet and Tillé (2006), where the inclusion probabilities are proportional to $x_k, k \in U$. To have a sampling design close to maximum entropy, a random sort of the population is performed before selection of the sample s. The covariance of the estimator of the mean is estimated using Formula (2.9). The cube algorithm is available in R in the *sampling* package, *samplecube* function, and a SAS macro is available on the INSEE website (Institut National de Statistique et des Études Économiques).

4. *SRSWOR sampling and MA estimator*

The estimator $\hat{\mu}_{MA}$ assisted by the model ξ is constructed using the auxiliary information given by $\mathbf{x}'_k = (1, x_k)$, where x_k is the mean consumption for the previous week. In these conditions, $\hat{\mu}_{MA}$ is the sum over any population U of the values $\hat{Y}_k$ estimated by the model (*cf.* Formula (2.13)). The covariance of the estimator of the mean is estimated using Formula (2.15).

4.2 Error of estimation of the mean curve

The error of estimation of the mean curve μ at instants $t_1, ..., t_{336}$, is evaluated according to the following criterion:

$$R_2\left(\hat{\mu}\right) = \frac{1}{336}\sum_{i=1}^{336}\left(\hat{\mu}\left(t_i\right) - \mu\left(t_i\right)\right)^2 \approx \frac{1}{T}\int_0^T \left(\hat{\mu}\left(t\right) - \mu\left(t\right)\right)^2 dt.$$

The results are shown in Table 4.3 for $I = 10,000$ simulations (replications). They clearly show that for this study, taking account of total consumption for the previous week substantially improves the precision of the estimate of the mean compared with simple random sampling without replacement, dividing the mean square error R_2 by 5. Among the different strategies, the best appear to be those that take account of auxiliary information via inclusion probabilities (STRAT, πps and systematic PPS).

Table 4.3
Square error R_2 of estimation of the mean μ, with $I = 10,000$ replications

Strategy	Mean	1st quartile	Median	3rd quartile
SRSWOR	40.53	10.82	22.16	51.09
STRAT (1)	5.78	3.68	5.08	7.07
STRAT (2)	6.49	4.03	5.48	7.88
πps	7.06	3.99	5.52	8.16
Systematic $\pi - ps$	6.73	3.85	5.20	8.07
MA	8.29	5.24	7.14	10.06

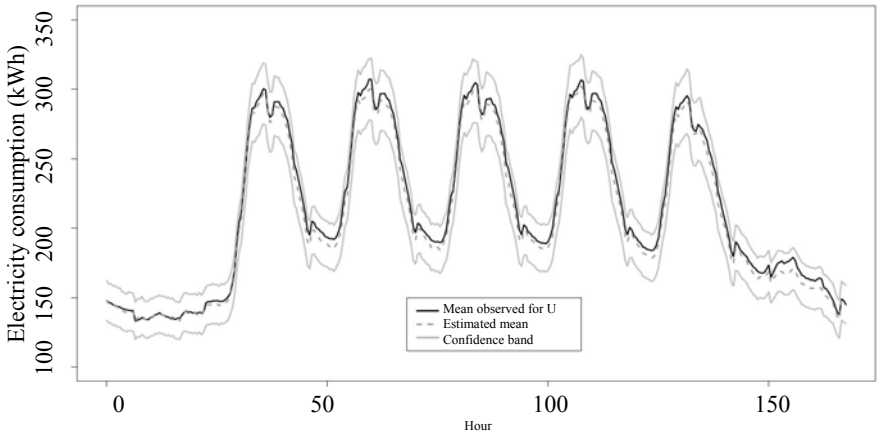
4.3 Coverage rate and width of confidence bands

The construction of confidence bands of level $1 - \alpha$ requires calculating quantiles of order $1 - \alpha$ of the supremum of Gaussian processes.

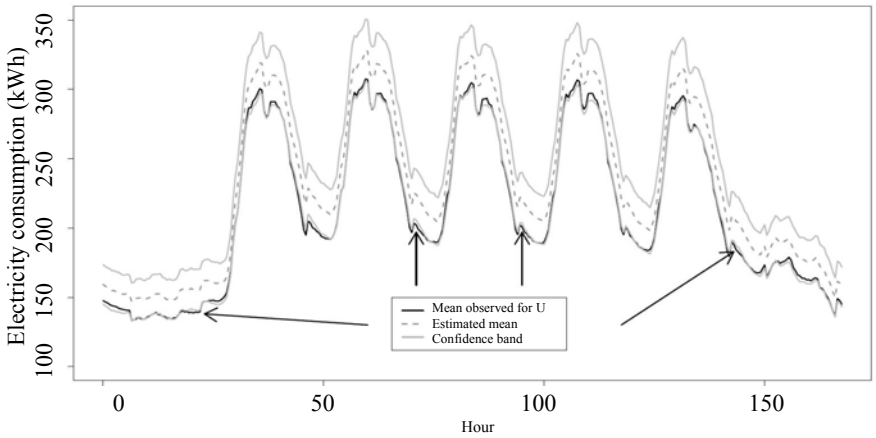
So as not to favour one method of constructing confidence bands over the other, we applied the two algorithms to the same sample s and we considered the same number M of processes. This number M varies from one estimator to the other owing to the computation time needed for the bootstrap approaches (see Section 4.4).

The empirical coverage rate is the proportion of times, among the $I = 2,000$ replications, where the true mean curve μ appears, for all instants t, within the confidence band constructed using an estimate $\hat{\mu}$. Figure 4.2 shows two examples of confidence bands (continuous grey curves) constructed from estimated curves (dotted grey curves). Figure 4.2(A) shows that the true mean curve for the population (continuous black curve) is within the confidence band at every instant. Conversely, Figure 4.2(B) shows that mean curve for the population is generally overestimated and there are a few instants (indicated by arrows) where the curve shown is outside the confidence band. Empirical coverage rates are shown in Table 4.4.

The two methods of constructing confidence bands yield coverage rates that are similar and fairly close to the desired nominal rates (95% and 99%). However, the results seem slightly less satisfactory for the πps designs and for the MA approach, for which the variance of the estimator is complex and more difficult to estimate precisely.



(A) The mean curve shown falls within the confidence band.



(B) The mean observed does not fall within the confidence band at the instants shown by the arrows.

Figure 4.2 Examples of confidence bands

Table 4.4
Empirical coverage rate (in %), for $I = 2,000$ replications

Method	Number M of processes	Bootstrap		Gaussian process	
		$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
SRSWOR	5,000	94.95	98.85	94.80	98.70
STRAT (1)	5,000	93.92	98.34	94.09	98.43
STRAT (2)	5,000	94.3	98.45	94	98.55
πps	1,000	94.73	98.77	93.87	98.61
MA	5,000	94.3	98.5	92.85	98.15

Another useful indicator is the mean width of the confidence band,

$$\frac{1}{336}\sum_{i=1}^{336} 2c_\alpha \hat{\sigma}\left(t_i\right) \approx \frac{1}{T}\int_0^T 2c_\alpha \hat{\sigma}\left(t\right) dt$$

the values of which are shown in Table 4.5. The two methods provide confidence bands of largely similar width. Also note that the use of the auxiliary variable considerably reduces the mean band width, which is cut in half if one of the stratified designs is used rather than a SRSWOR design.

Table 4.5
Mean width of confidence bands, for $I = 2,000$ replications

Method	Number M of processes	Bootstrap		Gaussian process	
		$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
SRSWOR	5,000	35.98	43.35	35.99	43.19
STRAT (1)	5,000	16.64	18.92	16.62	18.88
STRAT (2)	5,000	17.58	19.99	17.55	19.94
πps	1,000	17.85	20.31	17.62	19.93
MA	5,000	19.88	22.65	19.75	22.44

Figures 4.3 and 4.4 show the widths of the confidence bands for a level $\alpha = 0.05$, for each instant, depending on whether they are pointwise ($c_\alpha = 1.96$), estimated by simulations of Gaussian processes or obtained using the approach based on the Bonferroni inequality applied to each measurement point. We then have, in the latter case, $c_\alpha = 3.793048$, the quantile of order $1 - 0.05 / (336 \times 2)$ of a distribution $N(0,1)$. The bands obtained by Bonferroni are conservative, and they cover what might be considered the worst case in terms of information, the case of independence of the pointwise intervals. Note that the simulation approach substantially reduces the mean width of the bands in comparison with Bonferroni when the design does not allow all temporal information on the data to be taken into account (Figure 4.3). Conversely, for the stratified design (Figure 4.4), which provides a precise estimate of the mean curve, the confidence band constructed by simulation is close to that of Bonferroni, which can intuitively be interpreted as meaning that almost all the information was captured by the sampling design.

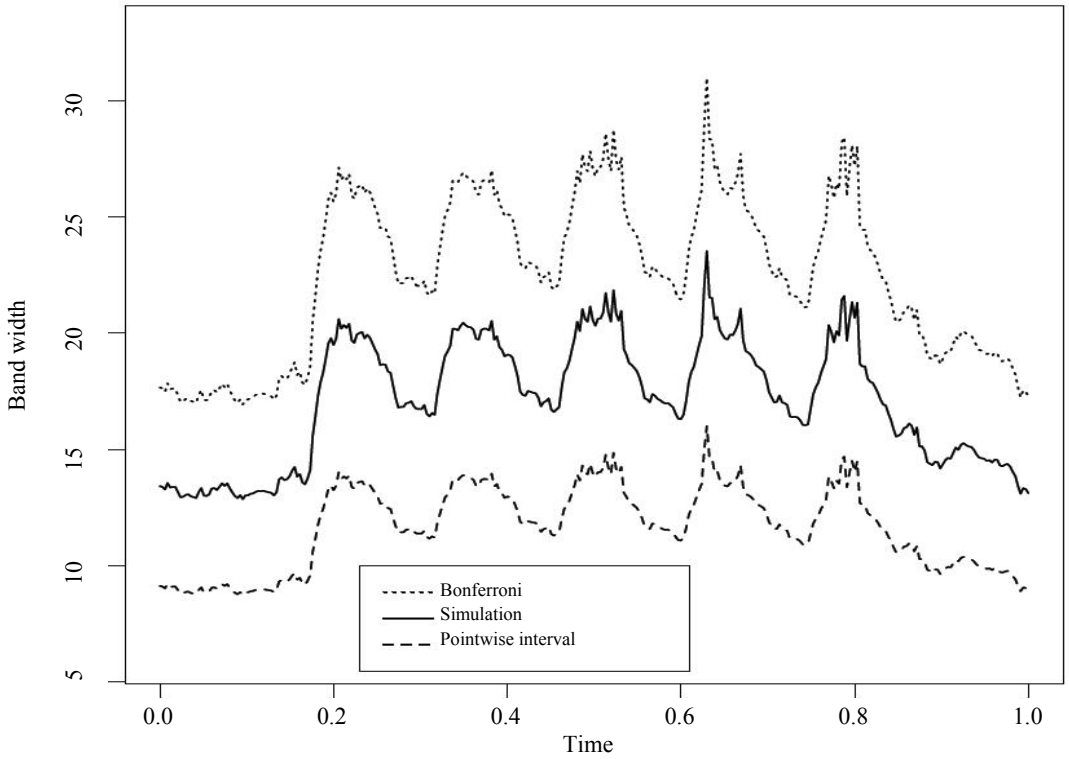


Figure 4.3 Simple random sampling without replacement. Width of confidence bands—pointwise, overall by process simulations, and with Bonferroni ($\alpha = 0.05$)

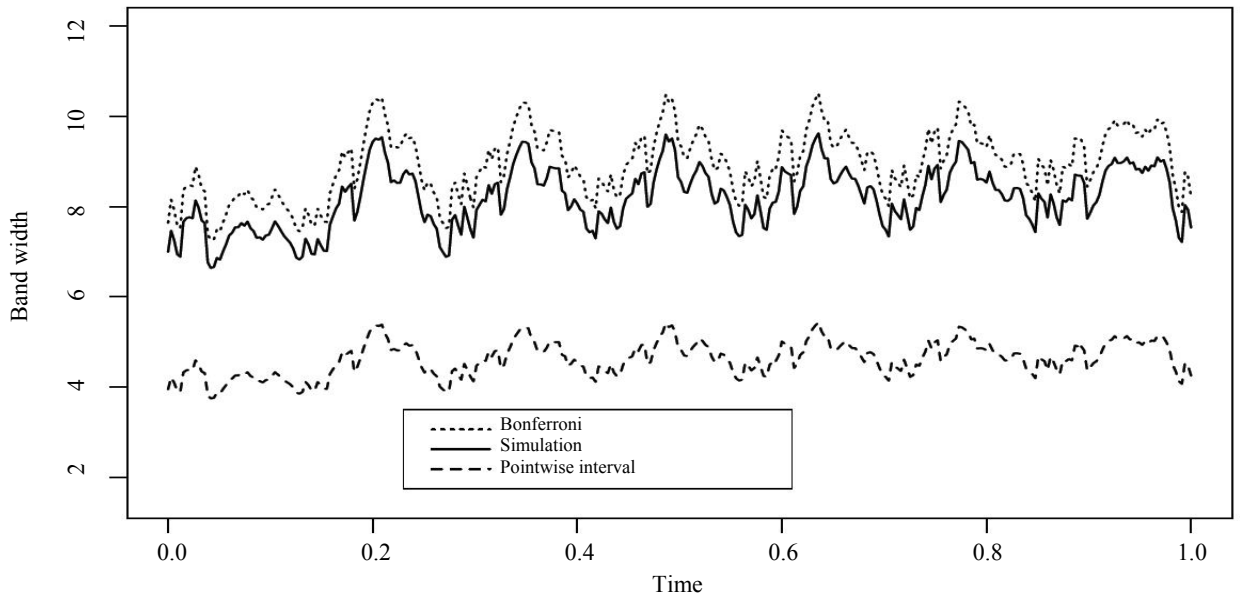


Figure 4.4 Stratified sampling (STRAT 1). Width of confidence bands—pointwise, overall by process simulations, and with Bonferroni (with $\alpha = 0.05$)

4.4 Computation time

Computation times with the bootstrap method are much greater—by a factor of approximately 1 to 1,000—than those with the Gaussian processes simulation method (*cf.* Table 4.6). The reason for this major difference is that in the bootstrap methods, the entire estimation process (construction of the fictitious population, drawing of a new sample, calculation of the estimator) must be repeated for each bootstrapped sample. Also, the designs that introduce auxiliary information are slower than SRSWOR, even though if used individually their computation time is entirely reasonable.

Table 4.6
Run time of a simulation in seconds for $M = 5,000$ replications. The SRSWOR, MA and STRAT strategies were programmed with R and πps with SAS.

Strategy	Bootstrap	Gaussian processes
SRSWOR	1,170.6	1.0
STRAT	1,839.5	1.4
πps	5,020.0	7.3
MA	3,156	1.4

5 Conclusion and perspectives for research

In this study, we have implemented and compared different strategies for using auxiliary information for estimating, and constructing confidence bands for, the mean of data in the form of curves. This information can be taken into consideration at the time of sampling by using unequal probability designs

or during estimation with simple random sampling without replacement, assisted by a functional-response regression model. It seems clear from our example of electricity consumption curves that when total consumption for the previous week is known, the precision of estimators of the mean can be greatly improved compared with an SRSWOR-type sampling.

Also, in this context of large samples and high-dimensional data, it also seems possible to construct, for these different strategies, confidence bands that have empirical coverage rates close to the desired rates. The two considered approaches—estimation of the covariance function and simulation of Gaussian or bootstrap processes—seem to perform comparably in terms of the width of the confidence bands; the main difference is in the computation time. The bootstrap, which seems more general because it does not require having a good estimator of the covariance function, proves to be much slower in practice.

Sometimes, in these flows of large-scale data, there are losses of information owing to signal transmission problems. The end result is that the utility has incomplete records of some trajectories. This issue, of partial non-response, can probably be dealt with by considering adaptations of classical non-response techniques (Haziza 2009) in the functional context. A fundamental question, then, is how to construct good estimators of the covariance function.

Acknowledgements

We wish to thank the anonymous referees as well as Guillaume Chauvet and Jean-Claude Deville for their helpful comments, which led to improvements in this study.

References

Bickel, P., and Krieger, A. (1989). Confidence bands for a distribution function using the bootstrap. *Journal of the American Statistical Association*, 84, 95-100.

Booth, J., Butler, R. and Hall, P. (1994). Bootstrap methods for finite population. *Journal of the American Statistical Association*, 89, 1282-1289.

Canty, A.J., and Davison, A.C. (1999). Resampling-based variance estimation for labour force surveys. *The Statistician*, 48, 379-391.

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

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

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

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 l1-median of a functional variable: Application to electricity load curves. *International Statistical Review*, 80, 40-59.

Chauvet, G. (2007). Méthodes de bootstrap en population finie. Ph.D. thesis, Université de Rennes II.

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

Cochran, W. (1977). Sampling techniques. New York: John Wiley & sons, Inc., 3rd Edition.

Cuevas, A., Febrero, M. and Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics and Data Analysis*, 51, 1063-1074.

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

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

Dessertaine, A. (2008). Estimation de courbes de consommation électrique à partir des mesures synchrones. In *Méthodes de Sondages* (Eds., P. Guibert, D. Haziza, A. Ruiz-Gazen and Y. Tillé), Dunod, France, 353-357.

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

Deville, J., and Tillé, Y. (2004). Efficient balanced sampling: The cube algorithm. *Biometrika*, 91, 893-912.

Deville, J., 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). *Oxford Handbook of Functional Data Analysis*. Oxford University Press.

Goga, C., and Ruiz-Gazen, A. (2013). Efficient estimation of nonlinear finite population parameters using nonparametrics, to appear in the *Journal of the Royal Statistical Society*, Series B, DOI: 10.1111/rssb.12024.

Gross, S. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*, 181-184.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In *Sample Surveys: Theory Methods and Inference*, volume 29 of Handbook of Statistics, (Eds., C. Rao and D. Pfeffermann), North-Holland, 215-246.

Madow, W. (1949). On the theory of systematic sampling, ii. *Annals of Mathematical Statistics*, 19, 535-545.

Ramsay, J., and Silverman, B. (2005). *Functional data analysis*. Springer, New York, second edition.

Rao, J., and Wu, C. (1988). Resampling inference with complex data. *Journal of the American Statistical Association*, 83, 231-241.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. Springer.

Sen, A. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5, 119-127.

Sitter, R.R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

Tillé, Y. (2011). Ten years of balanced sampling with the cube method: An appraisal. *Survey Methodology*, 37, 215-226.

Yates, F., and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B, 15, 235-261.