Estimation and confidence bands for the mean electricity consumption curve: a comparison of unequal probability sampling designs and model assisted approaches

Cardot, Hervé
Institut de Mathématiques de Bourgogne, UMR 5584 CNRS
9 Avenue Alain Savary
21078 Dijon, France
E-mail: herve.cardot@u-bourgogne.fr

Dessertaine, Alain
EDF, R&D, ICAME-SOAD
1 Avenue du Générale de Gaulle
92141 Clamart, France
E-mail: alain.dessertaine@edf.fr

Josserand, Etienne
Institut de Mathématiques de Bourgogne, UMR 5584 CNRS
9 Avenue Alain Savary
21078 Dijon, France
E-mail: etienne.josserand@u-bourgogne.fr

Lardin, Pauline
EDF, R&D, ICAME-SOAD
1 Avenue du Générale de Gaulle
92141 Clamart, France
E-mail: pauline.lardin@edf.fr

1 Introduction

We consider a survey sampling point of view in order to estimate the mean curve of large databases of functional data. When storage capacities are limited or transmission costs are high, selecting with survey techniques a small fraction of the observations is an interesting alternative to signal compression techniques. Our study is motivated, in such a context, by the estimation of the temporal evolution of mean electricity consumption curves. The French operator EDF has planned to install in a few years more than 30 millions electricity meters, in each firm and household, that will be able to send individual electricity consumptions at very fine time scales. Collecting, saving and analyzing all this information which can be seen as functional would be very expensive and survey sampling strategies are interesting to get accurate estimations at reasonable costs (Dessertaine, 2008). It is also well known that consumption profiles may depend on covariates such as past aggregated consumptions, meteorological characteristics (temperature, etc) or geographical information (altitude, latitude and longitude).

We compare in this work different ways of taking this information into account. A first one consists in using simple sampling designs, such as simple random sampling without replacement, and model assisted estimators (Särndal et al. 1992). A second strategy consists in considering unequal probability sampling designs such as stratified sampling or πps that can take additional information into account through their sampling weights.

The second question addressed in this work is how to build reliable confidence bands. When consistent estimators of the covariance function of the estimators are easy to build and the mean estimator satisfies a Functional Central Limit Theorem (Cardot and Josserand, 2011), a fast technique, inspired from Degras (2011), based on simulations of Gaussian processes in order to approximate the distribution of their suprema can be employed. This new approach is compared to bootstrap techniques which are also natural candidates for building confidence bands and that can be adapted to the finite population settings (Booth et al. 1994, Chauvet, 2007).

2 Functional data in a finite population

Let us consider a finite population $U = \{1, \ldots, k, \ldots, N\}$ of size N, and suppose we can observe, for each element k of the population U, a deterministic curve $Y_k = (Y_k(t))_{t\in[0,T]}$ that is supposed to belong to $C[0,T]$, the space of continuous functions defined on the closed interval $[0,T]$. Let us define the mean population curve $\mu \in C[0,T]$ by

$$(1)\quad \mu(t) \quad = \quad \frac{1}{N}\sum_{k\in U} Y_k(t), \quad t \in [0,T].$$

Consider now a sample s, *i.e.* a subset $s \subset U$, with known size n, chosen randomly according to a known probability distribution p defined on all the subsets of U. We suppose that all the individuals in the population can be selected, with probabilities that may be unequal, $\pi_k = \Pr(k \in s) > 0$ for all $k \in U$ and $\pi_{kl} = \Pr(k \;\&\; l \in s) > 0$ for all $k, l \in U$, $k \neq l$.

The Horvitz-Thompson estimator of the mean curve (Cardot et al. 2010), which is unbiased, is given by

$$(2)\quad \widehat{\mu}(t) \quad = \quad \frac{1}{N}\sum_{k\in s}\frac{Y_k(t)}{\pi_k} = \frac{1}{N}\sum_{k\in U}\frac{Y_k(t)}{\pi_k}\mathbb{1}_{k\in s}, \quad t \in [0,T].$$

In this context, we can define $\widehat{\mu}_{\text{srswor}}$, the simple random sampling without replacement mean estimator, by

$$(3)\quad \widehat{\mu}_{\text{srswor}}(t) \quad = \quad \frac{1}{n}\sum_{k\in s} Y_k(t), \quad t \in [0,T].$$

3 Estimators using auxiliary information

We consider now the particular case of stratified sampling with simple random sampling without replacement in all strata, assuming the population U is divided into a fixed number H of strata. This means that there is a partitioning of U into H subpopulations denoted by U_h, $(h = 1, \ldots, H)$. We can define the mean curve μ_h within each stratum h as $\mu_h(t) = N_h^{-1}\sum_{k\in U_h} Y_k(t),\ t \in [0,T]$, where N_h is the number of units in stratum h. The first and second order inclusion probabilities are explicitly known, and the mean curve estimator of $\mu_N(t)$ is

$$(4)\quad \widehat{\mu}_{\text{strat}}(t) \quad = \quad \frac{1}{N}\sum_{h=1}^{H} n_h^{-1} N_h \sum_{k\in s_h} Y_k(t),\ t \in [0,T],$$

where s_h is a sample of size n_h, with $n_h \leq N_h$, obtained by simple random sampling without replacement in stratum U_h.

Auxiliary information can be taken into account to build strata in order to improve the accuracy of the mean estimator. The sample size n_h in stratum h is determined by a Neyman-like allocation, as suggested in Cardot and Josserand (2011), in order to get a Horvitz-Thompson estimator of the mean trajectory whose variance is as small as possible.

Another interesting sampling design is the πps which can use directly auxiliary information. Indeed, we defined the first inclusion probability by

$$(5) \quad \pi_k = n\frac{x_k}{\sum_{k\in U} x_k},$$

where x_k is a real auxiliary variable for these k. For some units, π_k can be higher than one. To carry out this problem, we select automatically these units. Then, we compute again the first inclusion probabilities without the units already selected. We repeat this algorithm until all π_k are lower or equal to one. Using (2), we then obtain the πps mean estimator $\widehat{\mu}_{\pi\text{ps}}$.

Instead of using the auxiliary information into the sampling design, we can adjust a linear model and build a model assisted estimator $\widehat{\mu}_{\text{ma}}$. More precisely, we can write for all units k and $t \in [0,T]$

$$(6) \quad Y_k(t) = \beta_0(t) + \beta_1(t)x_k + \epsilon_{kt}$$

where $\beta_0(t)$ and $\beta_1(t)$ are regression coefficients (see Faraway, 1997). Survey sampling weights are taken into account to compute the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ of β_0 and β_1 (see Särndal et al. 1992). Finally, we get the mean estimator, for $t \in [0,T]$,

$$(7) \quad \widehat{\mu}_{\text{ma}}(t) = \frac{1}{N}\sum_{k\in s}\frac{Y_k(t)}{\pi_k} - \frac{1}{N}\left(\sum_{k\in s}\frac{\widehat{Y}_k(t)}{\pi_k} - \sum_{k\in U}\widehat{Y}_k(t)\right)$$

where $\widehat{Y}_k(t) = \widehat{\beta}_0(t) + \widehat{\beta}_1(t)x_k,\ t \in [0,T]$.

4 Confidence bands

In this section, we want to build confidence bands for μ of the form

$$(8) \quad \{[\widehat{\mu}(t) \pm c\,\widehat{\sigma}(t)]\,,\ t \in [0,T]\}\,,$$

where c is a suitable number and $\widehat{\sigma}(t)$ is an estimator of $\gamma(t,t)^{1/2}$, and where $\gamma(s,t) = \text{Var}\big(\widehat{\mu}(s), \widehat{\mu}(t)\big)$ is the covariance function of $\widehat{\mu}$. More precisely, given a confidence level $1-\alpha \in]0,1[$, we seek $c = c_\alpha$ that satisfies approximately

$$(9) \quad \mathbb{P}\left(\mu \in \{[\widehat{\mu}(t) \pm c_\alpha\,\widehat{\sigma}(t)]\,,\ \forall t \in [0,T]\}\right) = 1-\alpha.$$

4.1 Suprema of Gaussian processes

We consider the process $Z(t) = \big(\widehat{\mu}(t) - \mu(t)\big)/\widehat{\sigma}(t)$ which converges to a Gaussian process in the space of continuous functions $\mathcal{C}([0,T])$, under some technical assumptions (Cardot and Josserand, 2011). We can determine c_α such that

$$(10) \quad \mathbb{P}\left(|Z(t)| \le c_\alpha,\ \forall t \in [0,T]\right) = 1-\alpha,$$

where Z is a Gaussian process with mean zero and correlation function ρ, and where $\rho(s,t) = \widehat{\gamma}(s,t)/\big(\widehat{\gamma}(s,s)\ \widehat{\gamma}(t,t)\big)^{1/2}$. Note that the calculus of c_α with a Gaussian process is only possible when one can build an estimator $\widehat{\gamma}$ of the covariance function γ.

4.2 Bootstrap bands

Another way consists in estimating the covariance function by bootstrap (Booth et al. 1994, Chauvet, 2007). Using the sample s, we can generate a fictive population $U^\star$ and by simulation we obtain an approximation of $\widehat{\sigma}$. The following algorithm permits to build confidence bands:

1. Draw a sample s, with known size n, chosen randomly according to a known probability distribution p, and to compute $\widehat{\mu}$.
2. Duplicate each units $k \in s$ $1/\pi_k$ times to build a fictive population $U^\star$.
3. Draw in $U^\star$ M samples $s_j^\star$ with size n according to p, and to generate $\widehat{\mu}_j^\star(t)$, $j = 1, \dots, M$.
4. The function $\widehat{\sigma}(t)$ is estimated by the empirical standard deviation of $\widehat{\mu}_j^\star(t)$, $j = 1, \dots, M$.
5. Let $E_{c_\alpha} = \{j | \forall t \quad \widehat{\mu}(t) \in [\widehat{\mu}_j^\star(t) - c_\alpha \widehat{\sigma}(t); \widehat{\mu}_j^\star(t) + c_\alpha \widehat{\sigma}(t)]\}$. The coefficient c_α is chosen such that $\#(E_{c_\alpha}) = (1-\alpha)M$.

The second step of this algorithm may causes some problems because $1/\pi_k$ is not necessarily an integer. This is detailed in the next section and was already discussed by Booth *et al* (1994) and Chauvet (2007).

5 Study of mean electricity consumption curve

We consider now a population consisting in the $N = 15069$ electricity consumption curves measured during one week every half an hour. We have $d = 336$ time points. Note that our auxiliary information is the mean consumption, for each meter k, during previous week. We compare previous estimators with fixed size $n = 1500$. For each estimator we compute the confidence bands with the Gaussian bands and the bootstrap bands procedures. A draw back of Gaussian bands is that they require a covariance function estimator $\widehat{\gamma}$ whereas bootstrap methods just needs some adjustment to build a fictive population.

- $\widehat{\mu}_{\text{srswor}}$: For the simple random sampling without replacement estimator, we have an unbiased covariance function estimator

  (11)
  $$\widehat{\gamma}_{\text{srswor}}(s,t) = \left(\frac{1}{n} - \frac{1}{N}\right) \left(\frac{1}{n-1} \sum_{k,l \in s} Y_k(s) Y_l(t) - \frac{n}{n-1} \, \widehat{\mu}_{\text{srswor}}(s) \; \widehat{\mu}_{\text{srswor}}(t) \right), \quad s,t \in [0,T].$$

  To build the fictive population in the bootstrap step 2, we can remark that $1/\pi_k = N/n$ is not an integer. So, we duplicate $k \in s$ $[N/n]$ times, where $[.]$ is the entire part function. We complete the duplication step with a simple random sampling without replacement in s with a fixed size $N - n[N/n]$, in order to obtain a fictive population $U^\star$ which the size is equal to N.

- $\widehat{\mu}_{\text{strat}}$: The population is partitioned into $H = 10$ strata thanks to a k-means algorithm on our auxiliary variable, the mean consumption during the first week. The covariance function is estimated by

  $$(12) \quad \widehat{\gamma}_{\text{strat}}(s,t) = \frac{1}{N^2} \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} \, \widehat{\gamma}_h(s,t) \;\; s,t \in [0,T],$$

  where $\widehat{\gamma}_h$ is covariance function estimator into stratum h.

  The fictive population $U^\star$ is obtained by the same method used for $\widehat{\mu}_{\text{srswor}}$ in each stratum h.

- $\widehat{\mu}_{\pi\text{ps}}$: With the πps, it is difficult to obtain a formula for second inclusion probabilities because they depend on how the sample is drawn and there is no standard method. When the sample size is fixed and the sampling design p is close to the maximal entropy, we can use the Hajek formula (Berger, 1998) which can be adapted to approximate the covariance function

$$(13)\quad \widehat{\gamma}_{\pi\text{ps}}(s,t) = \frac{1}{N^2}\sum_{k\in s}(1-\pi_k)\left(\frac{Y_k(t)}{\pi_k}-\widehat{R}(t)\right)\left(\frac{Y_k(s)}{\pi_k}-\widehat{R}(s)\right)\quad s,t\in[0,T],$$

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

For the bootstrap, each $k\in s$ is duplicated $[1/\pi_k]$ times. As suggested in Chauvet (2007), to complete the population $U^\star$, we realize a πps sampling with an inclusion probability $\alpha_k = 1/\pi_k - [1/\pi_k]$. To keep a fixed sample size during the bootstrap step 3, we use the sampling design $p^\star$ defined for all $k\in U^\star$ by

$$(14)\quad \pi_k^\star = n\frac{x_k}{\sum_{k\in U^\star}x_k}.$$

- $\widehat{\mu}_{\text{ma}}$: The covariance function of the model assisted estimator is complicated to explicit because it depends on the sampling design and the model. By analogy with Breidt and Opsomer (2000), we have an asymptotic covariance estimator

$$(15)\quad \widehat{\gamma}_{\text{ma}}(s,t) = \frac{1}{N^2}\sum_{k,l\in s}\left(Y_k(s)-\widehat{Y}_k(s)\right)\left(Y_l(t)-\widehat{Y}_l(t)\right)\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l\pi_{kl}}\quad s,t\in[0,T],$$

where $\pi_k = \frac{n}{N}$ et $\pi_{kl}=\frac{n(n-1)}{N(N-1)}$ for $k,l\in s$ and $k\neq l$. To build the bootstrap bands, we adapt an algorithm of Helmers and Wegkamp (1998) to our case. For each sample s, we have $\widehat{\epsilon}_{kt} = \widehat{Y}_k(t)-\widehat{\beta}_0(t)-\widehat{\beta}_1(t)x_k$ for all $k\in s$. We then draw n *iid* realizations $Z_1,\dots,Z_n$ of a centered gaussian variable with unit variance and consider

$$(16)\quad Y_k^\star(t) = \widehat{\beta}_0(t)+\widehat{\beta}_1(t)x_k + Z_k\widehat{\epsilon}_{kt}\quad t\in[0,T].$$

By a simple random sampling without replacement in the fictive population $U^\star$, we obtain the mean estimator for $s^\star$

$$(17)\quad \widehat{\mu}^\star_{\text{ma}}(t) = \frac{1}{N}\sum_{k\in U}\tilde{Y}_k(t) - \frac{1}{N}\sum_{k\in s^\star}\frac{\tilde{Y}_k(t)-Y_k(t)}{\pi_k}\quad t\in[0,T].$$

where $\tilde{Y}_k(t)=\widehat{\beta}_0^\star(t)+\widehat{\beta}_1^\star(t)x_k$ for all $k\in U$, $\widehat{\beta}_0^\star$ and $\widehat{\beta}_1^\star$ are model parameters computing on $s^\star$, and $\pi_k = n/N$ for all $k\in s^\star$.

We will present in details the simulation results for these estimators during the talk. Briefly, we note that both methods employed to build the confidence bands give almost the same coverage and are close to nominal level of confidence. Moreover, confidence bands areas are very close too. The Gaussian processes simulation bands are much faster to compute but require a reliable estimator of the covariance function. On the other hand, the bootstrap bands can be long to generate but need nothing more than the basic estimator.

Acknowledgement. Etienne Josserand thanks the *Conseil Régional de Bourgogne, France* for its financial support (FABER PhD grant).

REFERENCES

Berger, Y. G. (1998). Rate of convergence for asymptotic variance of the Horvitz–Thompson estimator. *J. Statist. Planning and Inference*, **74**, 149-168.

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

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, **28**, 1026-1053.

Cardot, H., Chaouch, M., Goga, C. end C. Labrure (2010). Properties of Design-Based Functional Principal Components Analysis. *J. Statist. Planning and Inference*, **140**, 75-91.

Cardot, H., Josserand, E. (2011). Horvitz-Thompson Estimators for Functional Data: Asymptotic Confidence Bands and Optimal Allocation for Stratified Sampling. *Biometrika*, **98**, 107-118.

Chauvet, G. (2007). Méthodes de Bootstrap en population finie. *PhD Thesis*, Université Rennes II, France.

Degras, D. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica*, to appear.

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

Faraway, J. (1997). Regression Analysis for a Functional Responses. *Technometrics*, **39**, 254-261.

Helmers, R., Wegkamp, M. (1998). Wild Bootstrapping in Finite Population with Auxiliary Information. *Scandinavian journal of statistics*, **25**, 383-399.

Särndal, C.E., Swensson, B. and J. Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag.