# Semiparametric models with functional responses in a survey sampling setting : model assisted estimation of electricity consumption curve

Hervé Cardot, Etienne Josserand

# SEMIPARAMETRIC MODELS WITH FUNCTIONAL RESPONSES IN A SURVEY SAMPLING SETTING : MODEL ASSISTED ESTIMATION OF ELECTRICITY CONSUMPTION CURVE

Hervé Cardot & Etienne Josserand

*Institut de Mathématiques de Bourgogne, UMR 5584 CNRS,*
*Université de Bourgogne, 9, Av. A. Savary - B.P. 47 870, 21078 Dijon, France*

**Abstract**

This work adopts a survey sampling point of view when one has to estimate the mean curve of large databases of functional data. When storage capacities are limited selecting with survey techniques a small fraction of the observations is an interesting alternative to signal compression techniques. We propose here to take account of real or multivariate auxiliary information available at a low cost for the whole population, with semiparametric model assisted approaches, in order to improve the accuracy of Horvitz-Thompson estimators of the mean curve. We first estimate the functional principal components with a design based point of view in order to reduce the dimension of the signals and then propose semiparametric models to get estimations of the curves that are not observed. This technique is shown to be really effective on a real dataset of 18902 electricity meters measuring every half an hour electricity consumption over two weeks.

**Résumé**

Ce travail adopte une approche de type sondage quand le but est d'estimer une courbe moyenne d'une grande base de données de données fonctionnelles. Lorsque les capacités de stockage sont limitées, grâce aux techniques de sondage, une petite partie des observations est une alternative intéressante par rapport aux techniques de compression. Nous proposons ici de prendre en considération une information auxiliaire réelle ou multivariée obtenu à moindre coût sur la population toute entière, avec une approche semiparamétrique de type modèle assisté, dans le but d'améliorer les estimateurs d'Horvitz-Thompson de la courbe moyenne. D'abord, nous estimerons les composantes principales afin de réduire la dimension des signaux, et ensuite nous utiliserons des modèles semiparamétriques pour estimer les courbes qui n'ont pas été observées. Cette technique se montre vraiment efficace sur une base de données réelle de 18902 courbes de consommation électrique mesurée toutes les demi heures pendant deux semaines.

**Mots clés :** Analyse fonctionnelle, Sondages.

# 1 Introduction

With the development of distributed sensors one can have access of potentially huge databases of signals evolving along fine time scales. Collecting in an exhaustive way such data would require very high investments both for transmission of the signals through networks as well as for storage. As noted in Chiky and Hébrail (2009) survey sampling procedures on the sensors, which allow a trade off between limited storage capacities and accuracy of the data, can be relevant approaches compared to signal compression in order to get accurate approximations to simple estimates such as mean or total trajectories.

Our study is motivated, in such a context of distributed data streams, by the estimation of the temporal evolution of 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 analysing 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, 2006). It is well known that consumption profiles strongly depend on covariates such as past consumptions, meteorological characteristics (temperature, nebulosity, *etc*) or geographical information (altitude, latitude and longitude). Taking this information into account at an individual level (*i.e* for each electricity meter) is not trivial. One way to achieve this consists in reducing first the high dimension of the data by performing a functional principal components analysis in a survey sampling framework with a design based approach (Cardot *et al.*, 2010). It is then possible to build models, parametric or nonparametric, on the principal component scores in order to incorporate the auxiliary variables effects and correct our estimator with model assisted approaches (Särndal *et al.*, 1992). Note that this strategy based on modeling the principal components instead of the original signal has already been proposed, with a frequentist point of view, by Chiou *et al.* (2003) with singel index models and Müller and Yao (2008) with additive models.

We present in section 2 the Horvitz-Thomposon estimator of the mean consumption profile as well as the functional principal components analysis. We develop, in section 3, model assisted approaches based on statistical modeling of the principal components scores and derive an approximated variance that can be useful to build global confidence bands. Finally, we illustrate, in section 4, this methodology which allows to improve significantly more basic approaches on a population of 18000 electricity consumption curves measured every half an hour over one week.

# 2 Functional data in a finite population

Let us consider a finite population $U = \{1, \ldots, k, \ldots, N\}$ with size $N$, and suppose we can observe, for each element $k$ of the population $U$, a deterministic curve $Y_k = (Y_k(t))_{t\in[0,1]}$ that is supposed to belong to $L^2[0,1]$, the space of square integrable functions defined on

the closed interval $[0,1]$ equipped with its usual inner product $\langle \cdot, \cdot \rangle$ and norm denoted by $\| \cdot \|$. Let us define the mean population curve $\mu \in L^2[0,1]$ by

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

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, which is unbiased, is given by

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

As in Cardot *et al.* (2010) we would like to describe now the individual variations around the mean function in a functional space whose dimension is as small as possible according to a quadratic criterion. Let us consider a set of $q$ orthonormal functions of $L^2[0,1]$, $\phi_1, \ldots, \phi_q$, minimize, according to $\phi_1, \ldots, \phi_q$, the remainder $R(q)$ of the projection of the $Y_k$'s onto the space generated by these $q$ functions

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

with

$$R_{qk}(t) = Y_k(t) - \mu(t) - \sum_{j=1}^{q} \langle Y_k - \mu, \phi_j \rangle \phi_j(t), \quad t \in [0,1].$$

Introducing now the population covariance function $\gamma(s,t)$,

$$\gamma(s,t) = \frac{1}{N} \sum_{k \in U} \left( Y_k(t) - \mu(t) \right) \left( Y_k(s) - \mu(s) \right), \quad (s,t) \in [0,1] \times [0,1],$$

Cardot *et al.* (2010) have shown that $R(q)$ attains its minimum when $\phi_1, \ldots, \phi_q$ are the eigenfunctions of the covariance operator $\Gamma$ associated to the largest eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q \geq 0$,

$$\Gamma \phi_j(t) = \int_0^1 \gamma(s,t) \phi_j(s) ds = \lambda_j \phi_j(t), \quad t \in [0,1], j \geq 1.$$

When observing individuals from a sample $s$, a simple estimator of the covariance function

$$\widehat{\gamma}(s,t) = \frac{1}{N} \sum_{k \in s} \frac{1}{\pi_k} \left( Y_k(t) - \widehat{\mu}(t) \right) \left( Y_k(s) - \widehat{\mu}(s) \right) \ (s,t) \in [0,1] \times [0,1], \tag{3}$$

allows to derive directly estimators of the eigenvalues $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_q$ and the corresponding eigenfuctions $\widehat{\phi}_1, \ldots, \widehat{\phi}_q$.

# 3 Semiparametric estimation with auxiliary information

Suppose now we have access to $m$ auxiliary variables $X_1, \ldots, X_m$ that are supposed to linked to the individual curves $Y_k$ and we are able to observe these variables, at a low cost, for every individual $k$ in the population. Taking this additional information into account would certainly be helpful to improve the accuracy of the basic estimator $\widehat{\mu}$. Going back to the decomposition of the individual trajectories $Y_k$ on the eigenfunctions,

$$Y_k(t) = \mu(t) + \sum_{j=1}^{q} \langle Y_k - \mu, \phi_j \rangle \phi_j(t) + R_{qk}(t), \quad t \in [0,1],$$

and borrowing ideas from Chiou *et al.* (2003) and Müller and Yao (2008), an interesting approach consists in modeling the population principal components scores $\langle Y_k - \mu, \phi_j \rangle$ with respect to auxiliary variables at each level $j$ of the decomposition on the eigenfunctions,

$$\langle Y_k - \mu, \phi_j \rangle \approx f_j(x_{k1}, \ldots, x_{km})$$

where the regression function $f_j$ can be parametric or not and $(x_{k1}, \ldots, x_{km})$ is the vector of observations of the $m$ auxiliary variables for individual $k$.

It is possible to estimate the principal component scores $\widehat{C}_{kj} = \langle Y_k - \widehat{\mu}, \widehat{\phi}_j \rangle$, for $j = 1, \ldots, q$ and all $k \in s$ and then build a design based least squares estimator for the functions $f_j$

$$\widehat{f_j} \;=\; \arg\min_{g_j} \sum_{k \in s} \frac{1}{\pi_k} \left( \widehat{C}_{kj} - g_j(x_{k1}, \ldots, x_{km}) \right)^2, \tag{4}$$

in order to construct the following model-assisted estimator $\widehat{\mu}_X$ of $\mu$:

$$\widehat{\mu}_x(t) \;=\; \widehat{\mu}(t) - \frac{1}{N} \left( \sum_{k \in s} \frac{\widehat{Y}_k(t)}{\pi_k} - \sum_{k \in U} \widehat{Y}_k(t) \right) \tag{5}$$

where the predicted curves $\widehat{Y}_k$ are estimated for all the individuals of the population $U$ thanks to the $m$ auxiliary variables,

$$\widehat{Y}_k(t) = \widehat{\mu}(t) + \sum_{j=1}^{q} \widehat{f}_j(x_{k1}, \ldots, x_{km})\ \widehat{v}_j(t),\ t \in [0,1].$$

# 4 Application : estimation of electricity consumption curves

We have a population of $N = 18902$ electricity meters that are able to send electricity consumptions every half an hour over a period of two weeks, so that we have $d = 336$

time points. We are interested in estimating the mean consumption curve over the second week and we suppose that we know the mean consumption, $\bar{Y}_k = \frac{1}{336}\sum_{j=1}^{336} Y_k(t_j)$, for each meter $k$ of the population over the first week. This mean consumption will play the role of auxiliary information. Note that meteorological variables are not available in this preliminary study.

We first perform a simple random sampling without replacement (SRSWR) with fixed size of $n = 2000$ electricity meters over the second week order to get $\widehat{\mu}$ and perform the functional principal components analysis (FPCA).

To evaluate the accuracy of estimator (5) we made 500 replications of the following scheme

- Draw a sample of size $n = 2000$ in population $U$ with SRSWR and estimate $\widehat{\mu}, \widehat{\phi}_1$ and $\widehat{C}_{k1}$, for $k \in s$, over the second week.
- Estimate a linear relationship between $X_k$ and $\widehat{C}_{k1}$, for $k \in s$ where $X_k = \frac{1}{336}\sum_{j=1}^{336} Y_k(t_j)$ is the mean consumption over the first week, $\widehat{C}_{k1} \approx \widehat{\beta}_0 + \widehat{\beta}_1 X_k$.
- Estimate $\widehat{\mu}_X$ taking the auxiliary information into account with equation (5).

The following loss criterion $\int |\mu(t) - \widehat{\mu}(t)|dt$ has been considered to evaluate the accuracy of the estimators $\widehat{\mu}$ and $\widehat{\mu}_X$.

We will present the detail results during the presentation and we will show that model assisted estimators allow a significant improvement compared to the basic SRSWR approach.

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

## References

[1] CARDOT, H., CHAOUCH, M., GOGA, C. and C. LABRUÈRE (2010). Properties of Design-Based Functional Principal Components Analysis, *J. Statist. Planning and Inference.*, **140**, 75-91.

[2] CARDOT, H., JOSSERAND, E. (2009). Horvitz-Thompson Estimators for Functional Data: Asymptotic Confidence Bands and Optimal Allocation for Stratified Sampling. `http://arxiv.org/abs/0912.3891`.

[3] CHIKY, R., HEBRAIL, G. (2009). Spatio-temporal sampling of distributed data streams. *J. of Computing Science and Engineering*, to appear.

[4] CHIOU, J-M., MÜLLER, H.G. and WANG, J.L. (2003). Functional quasi-likelihood regression models with smooth random effects. *J.Roy. Statist. Soc., Ser. B*, **65**, 405-423.

[5] DESSERTAINE, A. (2006). Sondage et séries temporelles: une application pour la prévision de la consommation électrique. *38èmes Journées de Statistique*, Clamart, Juin 2006.

[6] MÜLLER, H-G., YAO, F. (2008). Functional Additive Model. *J. Am. Statist. Ass.* **103**, 1534-1544.
[7] SÄRNDAL, C.E., SWENSSON, B. and J. WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.
[8] SKINNER, C.J, HOLMES, D.J, SMITH, T.M.F (1986). The Effect of Sample Design on Principal Components Analysis. *J. Am. Statist. Ass.* **81**, 789-798.