Using complex surveys to estimate the L_1-median of a functional variable: Application to electricity load curves

Camelia Goga[*]
IMB, Université de Bourgogne, DIJON - France
camelia.goga@u-bourgogne.fr

April 11, 2013

Abstract

Mean profiles are widely used as indicators of the electricity consumption habits of customers. Currently, Électricité De France (EDF), estimates class load profiles by using point-wise mean function. Unfortunately, it is well known that the mean is highly sensitive to the presence of outliers, such as one or more consumers with unusually high-levels of consumption. In this paper, we propose an alternative to the mean profile: the L_1-median profile which is more robust. When dealing with large datasets of functional data (load curves for example), survey sampling approaches are useful for estimating the median profile and avoid storing all of the data. We propose here estimators of the median trajectory using several sampling strategies and estimators. A comparison between them is illustrated by means of a test population. We develop a stratification based on the linearized variable which substantially improves the accuracy of the estimator compared to simple random sampling without replacement. We suggest also an improved estimator that takes into account auxiliary information. Some potential areas for future research are also highlighted.

Key Words: Horvitz-Thompson estimator, k-means algorithm, poststratification, stratified sampling, substitution estimator, variance estimation.

1 Introduction

In the next few years, the French electricity company (EDF) intends to install over 30 million electricity smart meters, in every firm and household in France. These meters will be able to send individual electricity consumption measures on very fine time scales. The new smart electricity meters will provide accurate and up-to-date electricity consumption data. In view of this new setting, the relevant variables, such as the consumption curve, may be considered as realizations of functional variables depending on a continuous time index t that is in the $[0, \mathcal{T}]$ rather than as multivariate vectors. The amount of load data will be enormous when all or almost all customers have smart meters. Collecting, saving and analyzing all this information, would be very expensive. For example, if measures are taken every 10 minutes during one year and if we are interested in estimating the total electricity consumption for residential customers, the data storage is of about 100 terabytes.

Taking only a sample of meters reduces the amount of data storage and allows getting accurate estimates of quantities of interest such as the total or the mean consumption curve and even more, the L_1-median. In the presence of consumers with very high levels of electricity consumption, the L_1-median is a more robust indicator of the distribution of data than the mean consumption curve.

[*] This paper is based on work done in collaboration with Mohamed Chaouch (EDF R&D, Clamart, France) and published in the *International Statistical Review*, **80**, 40-59, 2012.

The paper is structured as follows: Section 2 gives a brief description of the L_1-median and the main results concerning its estimation with survey data. A weighted estimator is suggested for the median and a variance estimator is also proposed. Section 3 gives a comparison of the estimation of the L_1-median curve using several sampling designs and estimators.

2 Functional Median in a Survey Sampling Framework

Let us consider the finite population $U = \{1, \ldots, N\}$ of size N and a functional variable $\mathcal{Y}$ defined for each element k of the population $U : Y_k(t)$, for $t \in [0, \mathcal{T}]$, with $\mathcal{T} < \infty$. Let $< \cdot, \cdot >$ and, $|| \cdot ||$, be the inner product and the norm, respectively, defined on $L^2[0, \mathcal{T}]$. The median curve calculated from $Y_1, \ldots, Y_N$ is defined as (Chaudhuri, 1996 and Gervini, 2008):

$$m_N = \arg\min_{y \in L^2[0,\mathcal{T}]} \sum_{k=1}^{N} ||Y_k - y||. \qquad (1)$$

For $Y_1, \ldots, Y_N \in \mathbb{R}^d$, m_N defined by the relation (1) arises as a natural generalization of the well-known characterization of the univariate median (Koenker and Basset, 1978), $q = \arg\min_\theta \sum_{k=1}^N |Y_k - \theta|$, and it was called the *spatial median* by Brown (1983), the L_1-*median* by Small (1990) and the *geometric median* by Chaudhuri (1992). Weber (1909) obtained m_N as the solution to a *location problem* in which the $Y_1, \ldots, Y_N$ are the planar coordinates of N customers, who are served by a company that wants to find an optimal location for its warehouse.

Supposing that Y_k, for all $k = 1, \ldots, N$, are not concentrated on a line, the median exists and is unique (Kemperman, 1987). It is the solution of the following estimating equation:

$$\sum_{k=1}^{N} \frac{Y_k - y}{||Y_k - y||} = 0 \qquad (2)$$

provided that $m_N \neq Y_k$ for all $k = 1, \ldots, N$.

The median defined in this way is a global and central indicator of the distribution of the data. It has some other desirable properties and the reader is referred to Ilmonen *et al.* (2012) for a recent review of them. We plot in Figure 1 the mean curve versus the L_1-median for the test population of $N = 18902$ French companies. The electricity consumption was measured every 30 minutes during one week.

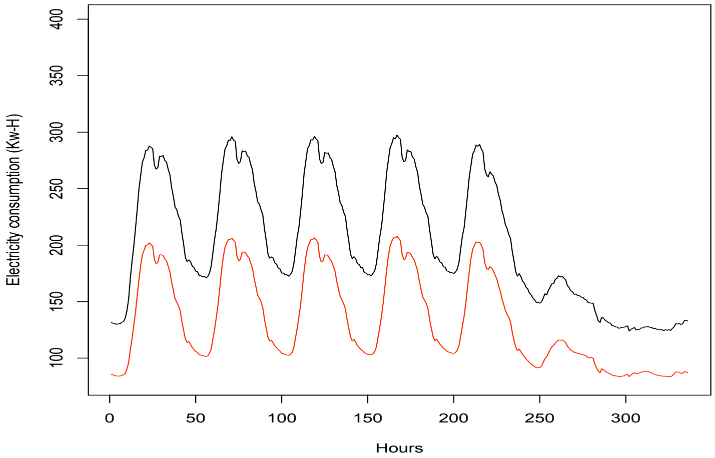


Figure 1: The spatial median profile is plotted in red and the mean profile in black.

2.1 The design-based estimator for the L_1 median

The median defined by (1) or (2) is computed by using iterative algorithms (Vardi and Zhang, 2000). These algorithms may be time-consuming especially if the size of the population is very large. In this work, we suggest estimating the median curve m_N by taking only a sample s from U according to a sampling design. Given a sample selection scheme, a probability measure $p(\cdot)$ on the set of all subsets of U, henceforth denoted $\mathcal{P}(U)$, is called a *sampling design.* A sample $s \subset \mathcal{P}(U)$ may be seen as the outcome of a random variable S whose probability distribution is specified by the function p. The subset of $\mathcal{P}(U)$ composed of those s for which $p(s)$ is strictly greater than zero constitutes the set of possible samples given the specified selection scheme.

For any $k \in U$, the probability that the unit k will be included in a sample is given by $\pi_k = \mathbb{P}(k \in S) = \sum_{k \in s} p(s)$ where the sum is considered over all samples s containing the individual k. If $k \neq l$ are two elements of U, the probability that both k and l are included in a sample is given by $\pi_{kl} = \mathbb{P}(k, l \in S) = \sum_{k,l \in s} p(s)$, where the sum is considered over all samples s containing both k and l.

The estimator $\widehat{m}_n$ of the L_1-median m_N is the solution of the minimization problem:

$$\widehat{m}_n = \arg\min_{y \in L^2[0,T]} \sum_{k \in s} \frac{||Y_k - y||}{\pi_k}. \tag{3}$$

Supposing now that $Y_k \neq \widehat{m}_n$ for all $k \in s$ and that Y_k are not concentrated on a line, we see following the same arguments as in Kemperman (1987) or Chaudhuri (1996), that the solution $\widehat{m}_n$ exists and is the unique solution of the design-based estimating equation:

$$\sum_{k \in s} \frac{1}{\pi_k} \frac{Y_k - \widehat{m}_n}{||Y_k - \widehat{m}_n||} = 0. \tag{4}$$

The median estimator $\widehat{m}_n$ is also called *the substitution estimator* of m_N and it is defined by a non-linear implicit function of Horvitz-Thompson estimators. As a consequence, the variance as well as the variance estimator of $\widehat{m}_n$ can not be obtained directly using Horvitz-Thomson formulas.

2.2 Asymptotic properties

Under broad assumptions, Chaouch and Goga (2012) give the following first-order expansion of $\widehat{m}_n$:

$$\widehat{m}_n \quad = \quad m_N + \sum_{k \in s} \frac{u_k}{\pi_k} - \sum_{k \in U} u_k + o_p(n^{-1/2}), \tag{5}$$

where $u_k = \Gamma^{-1}\left(\frac{Y_k - m_N}{||Y_k - m_N||}\right)$ and Γ is the Jacobian operator of the objective function from (1):

$$\Gamma = \sum_{k \in U} \frac{1}{||Y_k - m_N||}\left[\mathbf{I} - \frac{(Y_k - m_N) \otimes (Y_k - m_N)}{||Y_k - m_N||^2}\right],$$

with $\mathbf{I}$ the identity operator defined by $\mathbf{I}y = y$ and $\otimes$ the tensor product defined by $a \otimes b(y) =< a, y > b$. The quantity u_k for $k \in U$ is called *the linearized variable of* m_N and it is a kind of functional derivative. The linearized variable u_k for $k \in U$ is a curve and unknown.

From equation (5), we obtain that the median estimator $\widehat{m}_n$ may be approximated by $\sum_s \frac{u_k}{\pi_k}$ which is the Horvitz-Thompson estimator of $\sum_U u_k$. This is why estimating efficiently $\widehat{m}_n$ is equivalent to estimating efficiently the total of u_k. The asymptotic variance function of $\widehat{m}_n$ calculated under the sampling design is the variance of $\sum_{k \in s} u_k/\pi_k$, namely $\text{var}_p(\widehat{m}_n)(t) = \sum_{k \in U} \sum_{k \in U} (\pi_{kl} - \pi_k \pi_l) \frac{u_k(t)}{\pi_k} \frac{u_l(t)}{\pi_l}$. A variance estimator is given by $\widehat{\text{var}}_p(\hat{m}_n)(t) = \sum_{k \in s} \sum_{k \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{\hat{u}_k(t)}{\pi_k} \frac{\hat{u}_l(t)}{\pi_l}$ with $\hat{u}_k = \hat{\Gamma}^{-1}\left(\frac{Y_k - \widehat{m}_n}{||Y_k - \widehat{m}_n||}\right)$ and

$$\hat{\Gamma} = \sum_{k\in s} \frac{1}{\pi_k ||Y_k - m_N||} \left[\mathbf{I} - \frac{(Y_k - \hat{m}_n) \otimes (Y_k - \hat{m}_n)}{||Y_k - \hat{m}_n||^2} \right].$$

In practice, we observe the curves Y_k at D discretized points, $0 = t_1 \leq \ldots \leq t_D = \mathcal{T}$, which are supposed to be the same points for all $k \in U$. The curves may be seen then as multidimensional vectors, $\mathbf{Y}_k = (Y_k(t_1), \ldots, Y_k(t_D))'$ and $\hat{\boldsymbol{u}}_k = (\hat{u}_k(t_1), \ldots, \hat{u}_k(t_D))'$ observed for all $k \in s$. To compute $\hat{\boldsymbol{u}}_k$ for $k \in s$, one solves the $D \times n$ dimensional system

$$\widehat{\Gamma}(\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_n) = \left(\frac{\mathbf{Y}_1 - \hat{m}_n}{||\mathbf{Y}_1 - \hat{m}_n||}, \ldots, \frac{\mathbf{Y}_n - \hat{m}_n}{||\mathbf{Y}_n - \hat{m}_n||} \right)$$

where $\hat{\Gamma}$ is a $D \times D$ symmetric matrix obtained from the formula given above. Chaouch and Goga (2012) analyzed the behavior of $\widehat{\text{var}}_p(\widehat{m}_n)$ for different sampling designs.

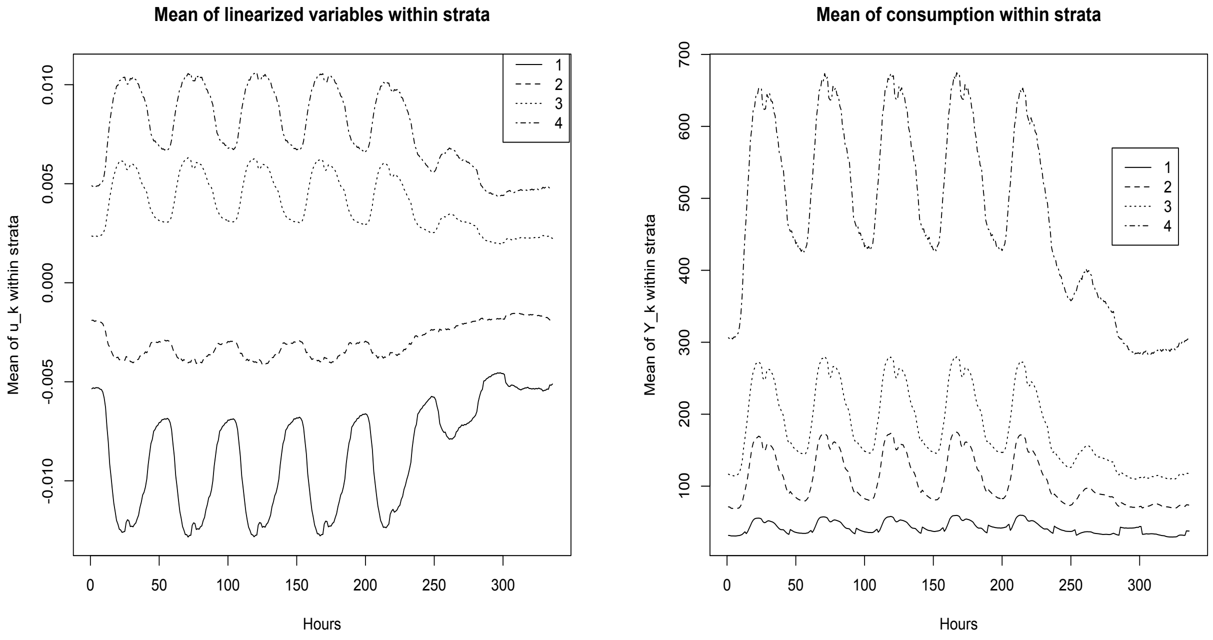
3 Application to the electricity load curves

General setting Let U be a population of $N = 18902$ electricity meters installed in small and large companies sending every 30 minutes the electricity consumption during a period of two weeks. We aim at estimating the median curve of the electricity consumption during the second week, using the consumption values recorded during the first week as auxiliary information. This means that we have measurements at 336 time points during each week. So, our study population of curves is a set of $N = 18902$ vectors $\mathbf{Y}'_k = (Y_k(t_1), \ldots, Y_k(t_D))$ with $D = 336$. Let X_k be the consumption curve for the kth firm recorded during the first week. The consumption curves present low peaks corresponding to night-time measurements and high peaks corresponding to middle of the day measurements. The electricity consumption decreases roughly around the 250th time measurement which corresponds to the beginning of the weekend. The mean and median curves present the same pattern as shown in Figure 1.

We consider several strategies of fixed size $n = 2000$ and compare them through simulations. We distinguish two kinds of sampling designs, based on whether they use or do not use auxiliary information. If auxiliary information is used at the sampling stage, some changes are needed because the variables involved now are curves. In the opposite situation, the selection of the sample is realized from the sampling frame list as for classical multivariate surveys. Finally, the frame list of French firms is well-constructed being very often updated and most of the designs considered below are often used in practice.

Simple random sampling without replacement (SRSWOR) of size $n = 2000$ from the population U of size N consists of taking n elementss from the list of N companies. For each selected company, we record its consumption electricity of each time point. The median estimator $\widehat{m}_n$ is obtained from equation (4) with $\pi_k = n/N$ for all $k \in U$.

Stratified sampling (STRAT) In this case, U is split into H non-overlapping sub-populations U_h of size N_h, $h = 1 \ldots, H$. The consumption electricity during the first week is used to construct $H = 4$ strata homogeneous with respect to $\mathbf{Y}$. We select a SRSWOR sample s_h of size n_h from each stratum U_h, $h = 1 \ldots, H$. The proportional and the optimal allocations $n_h, h = 1, \ldots, H$ are used. $\hat{m}_n$ is obtained from equation (4) with $\pi_k = n_h/N_h$ for all $k \in U_h$. STRAT sampling is very efficient if the strata are homogeneous with respect to the linearized variable. We plot in Figure 3 (b), the consumption mean within strata during the second week. We notice that stratum 4 corresponds to consumers with high global levels of consumption, whereas stratum 1, corresponds to consumers with low global consumption. Figure 3 (a) gives the mean curves of the linearized variable within strata and computed for the second week. As with the first stratification, the population of the linearized variable curves is also stratified.



(a)

(b)

Figure 2: Stratification based on the consumption curve: (a) Mean of linearized variables u_k within each stratum. (b) Mean of the consumption curve Y_k within each stratum

Proportional to size sampling without replacement (πPS) In πPS sampling, the sampling is without-replacement and the probability π_k that the individual k belongs to a sample is proportional to the mean of $X_k(t)$ over all $t = 1, \ldots, D$ where D is the number of discretization points in the interval $[0, \mathcal{T}]$. This means that $\pi_k = n\tilde{X}_k / \sum_{k \in U} \tilde{X}_k$, where $\tilde{X}_k = \sum_{t=1}^{D} X_k(t)/D$. The Horvitz-Thompson estimator of the median is obtained by solving the equation (Lardin, Cardot and Goga, 2013):

$$\sum_{k \in s} \frac{Y_k - \widehat{m}_n}{\pi_k ||Y_k - \widehat{m}_n||} = 0. \tag{6}$$

SRSWOR with Poststratification (POST) We split the population U into $G = 4$ *post-strata* according to values of the linearized variable u_k computed during the first week. This means that the post-strata are homogeneous with respect to the linearized variable. Nevertheless, we do not use this partitioning of U to conduct stratified sampling. We select a SRSWOR sample of size n and for each sampled unit k, we determine to which post-stratum it belongs. The group membership totals N_g are known for all $g = 1, \ldots, G$ and this auxiliary information may be used to get an improved estimator of m_N. Let $s_g = s \cap U_g$. Then, the poststratified median estimator satisfies

$$\sum_{g=1}^{G} \sum_{k \in s_g} \frac{N_g}{n_g} \frac{Y_k - \widehat{m}_n}{||Y_k - \widehat{m}_n||} = 0. \tag{7}$$

Chaouch and Goga (2012) compared these designs and estimators through 500 simulations and by using the following loss criteria:

$$R(\hat{m}_n) = \int_0^{\mathcal{T}} |\hat{m}_n(t) - m_N(t)| \simeq \frac{1}{D} \sum_{d=1}^{D} |\hat{m}_n(t_d) - m_N(t_d)|$$

We observed that clustering the space of functions by performing stratified sampling leads to an important gain compared to simple random sampling without replacement. We note that the poststratification gives better results than those obtained with stratified sampling because the post-strata are homogeneous with respect to the linearized variable u_k while the strata are homogeneous with respect to the consumption electricity Y_k. STRAT with proportional allocation gives slightly better results than those obtained with the optimal allocation. This is due to the fact that the optimal allocation is computed by minimizing the variance for the mean estimator; we are, however, interested here in estimating the median curve. Finally, the πPS sampling performs rather poorly for the estimation of the median. Lardin, Cardot and Goga (2013) suggest B-spline smoothing of the sampling weights $1/\pi_k$ and this improves greatly the performance of the πPS design.

	Mean	1^{st} quartile	median	3^{rd} quartile
SRSWOR	2.531	1.322	1.982	3.351
STRAT+PROP	1.7370	1.0470	1.4860	2.2480
STRAT+OPT	2.2940	1.4660	1.9790	2.7830
πPS	7.399	2.869	6.050	10.480
POST	1.041	0.8275	0.9785	1.203

Table 1: Estimation errors for m_N.

References

Brown, B.M. (1983). Statistical Use of the Spatial Median, *Journal of the Royal Statistical Society*, B, **45**, 25-30.

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**, 40-59.

Chaudhuri, P. (1996). On a Geometric Notion of Quantiles for Multivariate Data, *Journal of the American Statistical Association*, **91**, 862-872.

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

Ilmonen, P., Oja, H. and Serfling, R. (2012). On Invariant Coordinate System (ICS) Functionals. *International Statistical Review*, **80**, 93-110.

Kemperman, J.H.B. (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*, 217-230.

Koenker, R., and Bassett, G. (1978). Regression Quantiles, *Econometrica*, **46**, 33-50.

Lardin, P., Cardot, H. and Goga, C. (2013). Analyzing large datasets of functional data: a survey sampling point of view, *preprint*.

Small, C.G. (1990). A survey of multidimensional medians, *International Statistical Review*, **58**, 263-277.

Vardi, Y and Zhang, C.H. (2000). The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA*, **97**, 1423-1426.

Weber, A. (1909). Uber Den Standard Der Industrien, Tubingen. English translation by C. J. Freidrich (1929), in Alfred Weber's Theory of Location of Industries, Chicago: Chicago University Press.