Estimation d'une courbe moyenne de consommation électrique par sondage en présence de valeurs manquantes

Anne De Moliner [1,2] & Hervé Cardot [2] & Camelia Goga [2]

[1] *EDF R&D, Clamart, anne.de-moliner@edf.fr*
[2] *Institut de Mathématiques de Bourgogne UMR CNRS 5584, Université de Bourgogne, Dijon, France.* {*herve.cardot,camelia.goga*}*@u-bourgogne.fr*

Résumé. Dans un futur proche, des dizaines de millions de courbes de charges (i.e. consommations d'électricité mesurées à un pas de temps fin, ici demi horaire) de ménages français seront disponibles. Ces données constitueront une masse d'information considérable, qui pourrait être exploitée grâce à des techniques d'échantillonnage, afin d'estimer par exemple la consommation totale de différents segments de clientèle ou périmètres de fournisseurs. Malheureusement différents aléas techniques pourraient générer des valeurs manquantes qui risqueraient de détériorer la précision des estimateurs voire de créer des biais. Dans cette communication, nous proposons donc une méthode d'estimation de courbe de consommation moyenne en présence de données manquantes basée sur l'extension de méthodes d'estimation non paramétrique sur données fonctionnelles observées en peu d'instants de mesure (Hall *et al.* (2006)) ou longitudinales (Staniswalis & Lee (1998)) au cas où les courbes sont collectées par sondage. Une approximation de variance basée sur la linéarisation (inspirée de Särndal *et al.* 1992) pour la courbe moyenne estimée est également proposée. Notre méthode est comparée avec les techniques préexistantes pour différents types d'échantillonnage et différents scénarios de valeurs manquantes sur des jeux de données réels.

Mots-clés. Données fonctionnelles, estimateur à noyau, estimation de variance, industrie, valeurs manquantes.

Abstract. In the near future, tens of millions of load curves measuring the electricity consumption of French households in small time intervals (probably half hours) will be available. All these collected load curves represent a huge amount of information which could be exploited using sampling techniques. In particular, the total consumption of a specific customer group (for example all the customers of an electricity supplier) could be estimated using random sampling methods. Unfortunately, data collection may undergo technical problems resulting in missing values. The aim of this communication is to present a new estimation method for the mean curve in the presence of missing values. Our method consists in extending nonparametric techniques for sparse functional data analysis (Hall *et al.* 2006) or longitudinal data analysis (Staniswalis & Lee, 1998) to curves collected by sampling. We also propose a variance estimator (inspired by Särndal *et al.* 1992) for the estimated mean load curve. Then we compare this new method

to preexisting ones on real datasets for different sampling designs and missing values patterns.

Keywords. Functional data, industry, missing values, kernel smoothers, variance estimation.

1 Contexte et problématique

La quantité d'information disponible pour le fournisseur et le distributeur d'énergie va connaître une croissance fulgurante dans les prochaines années. En particulier, des dizaines de millions de courbes de charge, c'est-à-dire de séries de consommations à un pas de temps fin, probablement demi-horaire, d'entreprises et de ménages français seront disponibles. Le stockage et l'exploitation de données massives constituant une problématique complexe, il serait envisageable d'utiliser des techniques d'échantillonnage afin de reconstituer des consommations agrégées, appelées synchrones de consommations, au niveau d'un périmètre particulier (fournisseur, segment marketing, équipement particulier,...).

Comme tout processus industriel de masse, la collecte des données est susceptible de subir toutes sortes d'aléas techniques le long de la chaîne de mesure et de remontée d'information. Les données pourraient ainsi contenir des valeurs manquantes. Ce problème s'apparente à celui de la non réponse dans les enquêtes par sondages : il détériore la précision des estimateurs et peut éventuellement créer des biais si le mécanisme de défaillance n'est pas indépendant des valeurs mesurées. L'estimation en présence de valeurs manquantes fait l'objet d'une abondante littérature (voir par exemple Haziza (2009)) mais à notre connaissance le cas où les données collectées sont des courbes n'a pas été traité.

On se propose donc ici d'utiliser des méthodes d'analyse de données fonctionnelles en les adaptant au cadre des sondages ainsi qu'à la présence de données manquantes afin d'exploiter au mieux les spécificités de notre problème, à savoir les fortes corrélations entre les consommations aux différents instants ainsi que la régularité des courbes. Plus précisément, on partira de l'estimateur non paramétrique proposé par Staniswalis & Lee (1998).

2 L'estimateur de la courbe moyenne

On considère une population U constituée de N clients. A chacun de ces clients k on associe une trajectoire (la courbe de charge) $Y_k(t)$. Chaque courbe est mesurée en un ensemble p d'instants de mesure équidistants (un par demi-heure par exemple) au cours de la période $[0, T]$: $0 \leq t_1 < \ldots < t_j < \ldots < t_p \leq T$ et l'objectif est d'estimer la courbe de charge moyenne dans la population :

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

Si on connaissait l'ensemble des trajectoires de la population en $t_1, \ldots, t_p$, on pourrait utiliser l'estimateur à noyau suivant (Staniswalis & Lee, 1998) :

$$\tilde{\mu}(t) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{p} K(h^{-1}(t-t_j))Y_i(t_j)}{\sum_{i=1}^{N}\sum_{j=1}^{p} K(h^{-1}(t-t_j))} = \sum_{j=1}^{p} w(t,t_j,h)\mu(t_j),$$

pour estimer en chaque instant $t \in [0,T]$ la trajectoire moyenne, où K est un noyau, contrôlé par une fenêtre h, et les poids $w(t,t_j,h)$ ne dépendent que de K, de h et des instants t_j. On peut de plus montrer, sous certaines hypothèses de régularité sur les trajectoires et si la fenêtre h n'est pas trop grande, que l'erreur d'approximation de $\mu(t)$ par $\tilde{\mu}(t)$ est négligeable devant l'erreur due à l'échantillonnage (Cardot *et al.* 2013).

Il arrive que tout ou une partie de la courbe de certains individus n'est pas observée du fait de défaillances techniques. On introduit alors un processus de non réponse $r_k(t)$ qui vaut 1 si la donnée est présente pour l'individu k à l'instant t et 0 sinon. Il sera supposé aléatoire et indépendant de la quantité mesurée. On note $\theta_k(t_j) = \Pr(r_k(t_j)=1)$ et $\theta_k(t_j,t'_j) = \Pr(r_k(t_j)=1 \;\&\; r_k(t'_j)=1)$. Un échantillon s de taille n est tiré dans la population U selon un plan de sondage et on note $\pi_k = \Pr(k \in s)$ et $\pi_{kl} = \Pr(k \in s \& l \in s)$.

Adapté à notre contexte (en rajoutant les poids de sondage $1/\pi_k$, les indicatrices et les probabilités de non réponse), un estimateur de μ en t est donné par

$$\widehat{\mu}_r(t) = \frac{\sum_{j=1}^{p} K(h^{-1}(t-t_j)) \left(\sum_{k\in s} \frac{r_k(t_j)}{\theta_k(t_j)} \frac{Y_k(t_j)}{\pi_k} \right)}{\sum_{j=1}^{p} K(h^{-1}(t-t_j)) \left(\sum_{k\in s} \frac{r_k(t_j)}{\theta_k(t_j)} \frac{1}{\pi_k} \right)}. \tag{1}$$

3 Calcul et estimation de la variance

On souhaite pouvoir fournir un intervalle de confiance associé à notre estimateur sans employer de techniques de rééchantillonnage, trop coûteuses en temps de calcul, et on va donc proposer une approximation de la variance de $\widehat{\mu}_r(t)$ basée sur une technique de linéarisation inspirée par Särndal (1992), Chapitre 15.

Notre estimateur $\widehat{\mu}_r(t)$ s'écrit comme un ratio:

$$\widehat{\mu}_r(t) = \frac{\sum_{k\in s} \frac{1}{\pi_k} \sum_{j=1}^{p} \frac{\check{Y}_{kj}(t)}{\theta_k(t_j)} r_k(t_j)}{\sum_{k\in s} \frac{1}{\pi_k} \sum_{j=1}^{p} \frac{\check{X}_j(t)}{\theta_k(t_j)} r_k(t_j)} = \frac{\widehat{Y}}{\widehat{X}} \tag{2}$$

où $\check{Y}_{kj}(t) = K(h^{-1}(t-t_j))Y_k(t_j)$ et $\check{X}_j(t) = K(h^{-1}(t-t_j))$. Notons s_r le sous-échantillon des répondants. En utilisant la décomposition classique de la variance, on a

$$V(\widehat{Y}(t)) \;=\; V_R E_p(\widehat{Y}(t)|s_r) + E_R V_p(\widehat{Y}(t)|s_r)$$

où E_p (resp. V_p) est l'espérance (resp. la variance) par rapport au plan de sondage et E_R (resp. V_R) est l'espérance (resp. la variance) par rapport au mécanisme de non réponse. En déroulant les calculs pour chacun des deux termes on trouve finalement que :

$$\begin{aligned} V(\widehat{Y}(t)) &= \sum_{k\in U}\sum_{l\in U}\frac{\Delta_{kl}}{\pi_k\pi_l}\widetilde{Y}_k(t)\widetilde{Y}_l(t) + \sum_{k\in U}\frac{1}{\pi_k}\sum_{j,j'=1}^{p}\frac{\check{Y}_{kj}(t)\check{Y}_{kj'}(t)}{\theta_k(t_j)\theta_k(t_{j'})}(\theta_k(t_j,t_{j'}) - \theta_k(t_j)\theta_k(t_{j'})) \\ &= V_1 + V_2 \end{aligned} \tag{3}$$

où $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$, $\widetilde{Y}_k(t) = \sum_{j=1}^{p}\check{Y}_{kj}(t)$ et $\widetilde{Y}_{k,r}(t) = \sum_{j=1}^{p}\dfrac{\check{Y}_{kj}(t)}{\hat{\theta}_k(t_j)}r_k(t_j)$. Le terme V_1 est la variance due à l'échantillonnage (lorsqu'il n'y a pas de non-réponse) et le terme V_2 est la variance due à la non-réponse.

Ensuite, en remarquant que notre estimateur final est un ratio de totaux et en utilisant la technique de la linéarisation (Deville, 1999), on en déduit l'approximation suivante de la variance :

$$V(\widehat{\mu}_r(t)-\tilde{\mu}(t)) \simeq \sum_{k\in U}\sum_{l\in U}\frac{\Delta_{kl}}{\pi_k\pi_l}\widetilde{u}_k(t)\widetilde{u}_l(t) + \sum_{k\in U}\frac{1}{\pi_k}\sum_{j,j'=1}^{p}\frac{u_{kj}(t)u_{kj'}(t)}{\theta_k(t_j)\theta_k(t_{j'})}(\theta_k(t_j,t_{j'})-\theta_k(t_j)\theta_k(t_{j'}))$$

où $u_{kj}(t) = \dfrac{1}{\sum_{k\in U}\sum_{j=1}^{p}\check{X}_j(t)}\left(\check{Y}_{kj}(t) - \tilde{\mu}(t)\check{X}_j(t)\right)$ et $\widetilde{u}_k(t) = \sum_{j=1}^{p}u_{kj}(t)$.

L'estimateur final de la variance, pour un plan de sondage quelconque, s'obtient en remplaçant les totaux présents dans les linéarisées par les estimateurs de Horvitz-Thompson. Les probabilités de réponse, si elle ne sont pas connues, peuvent également être estimées sous des hypothèses simplificatrices sur le mécanisme de non réponse : indépendant de l'individu, constant au cours du temps, stationnaire, *etc*

4 Application : estimation de courbes de charge

On dispose de 20 000 courbes de charge de clients pendant une période d'un mois ainsi que d'informations auxilaires (tarif du client et consommation de l'année précédente). Sur ce jeu de données, on va simuler différents tirages d'échantillons avec valeurs manquantes, pour des scénarios d'échantillonnage (sondage aléatoire simple, **sondage aléatoire stratifié selon la consommation**, et sondage aléatoire stratifié selon la consommation et un critère de forme) et de non réponse variés (trous non simultanés d'une journée, d'une semaine, de moins de 3h, trous simultanés, **combinaison de ces différentes scénarios**,

pour **10%** ou 20% de valeurs manquantes). Le paramétrage en gras correspondra à notre scénario de référence.

Pour ces différents scénarios, on tirera 100 échantillons de 2000 individus sur lesquels on simulera de la non réponse, puis on comparera les performances de notre méthode à celles de méthodes de référence (imputation à la moyenne de la strate, recopie de la valeur sept jours avant, et enfin interpolation linéaire des trous de moins de trois heures et imputation à la moyenne de classe au-delà). On regardera le critère MAPE qui est la moyenne, sur les échantillons et les instants, de l'écart relatif entre la courbe estimée et la vraie courbe.

A un premier estimateur (estimateur "direct") construit en supposant que la probabilité de réponse $\theta_k(t_j)$ est constante : elle ne dépend ni de l'individu k ni de l'instant t_j, on va en fait préférer une alternative (estimateur "stratifié") qui consiste à mener l'estimation séparément strate par strate puis ensuite à agréger le tout en pondérant l'estimateur de chaque strate par son effectif. Cela permettra en particulier de pouvoir postuler un mécanisme de réponse différent pour chaque strate, ce qui apparaît comme une hypothèse plus réaliste. Le choix de la fenêtre h est crucial et doit permettre de minimiser l'erreur quadratique en trouvant un compromis entre biais (h élevé) et variance (h faible). Ce paramètre de lissage est choisi en utilisant une validation-croisée qui tient compte des poids de sondage (Cardot *et al.* 2013).

Pour le scénario de référence, les résultats sont les suivants :

IMPUTATION	Interpolation	J-7	Moyenne	Noyau Direct	Noyau Stratifié	Données complètes
MAPE	**1,007**%	7,532%	1,053%	1,233%	1,103%	**0.93%**

Table 1: Performances des méthodes d'imputation, Scénario de réference.

On constate que notre méthode donne des résultats corrects, comparables à ceux de l'imputation à la moyenne de classe. Néanmoins, la méthode la plus performante reste l'interpolation des petits trous et la moyenne de classe pour les gros trous. La recopie de la valeur à J-7 donne de mauvais résultats du fait de la non stationnarité des consommations pour la période observée. Enfin comme on le supposait, la version strate par strate de l'estimateur à noyau est plus performante que la version globale.

Les autres tests fournissent des résultats concordants et en particulier le plan de sondage ne semble pas avoir d'influence sur le classement relatif des différentes méthodes. La modification du scénario de non réponse permet quant à elle d'affirmer que l'interpolation linéaire fonctionne particulièrement bien sur les petits trous (moins de 3h) alors que la longueur de la série de valeurs manquantes n'a pas d'effet sur les performances des autres mthodes.

Enfin, un inconvénient important de notre nouvelle méthode se situe au niveau des temps de calcul : alors qu'un test (tirage de l'échantillon, simulation de la non réponse

et estimation) prend entre 2 et 10 minutes pour les techniques de référence, il prend 12h pour l'estimateur à noyau lorsque la fenêtre h est choisie par validation croisée.

5 Conclusions

Nous avons proposé ici une méthode d'estimation de courbe moyenne à partir d'un échantillon de courbes incomplètes, basée sur l'extension d'une méthode non paramétrique fonctionnelle au cadre des sondages avec données manquantes. La variance de cet estimateur peut être approximée par linéarisation.

Des tests sur des données réelles ont montré que la méthode donnait des résultats corrects, similaires à l'imputation par moyenne de classe. Néanmoins pour les trous de petite taille (moins de 3h), la meilleure méthode reste l'imputation par interpolation linéaire courbe à courbe. On peut cependant penser que notre estimateur, basé sur un lissage, aurait un intérêt comparatif par rapport aux autres lorsque les données d'entrée sont bruitées ou mesurées avec erreur.

Ce travail se poursuivra par le développement et le test d'une autre méthode d'estimation pour des échantillons de courbes de charge avec données manquantes, cette fois basée sur l'imputation par plus proche voisin.

Bibliographie

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

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

[3] Hall P., Mueller H.G. and Wang J.L. (2006). Properties of principal components methods for functional and longitudinal data analysis, *Annals of Statistics*, **34**, 1493-1517.

[4] Haziza, D. (2009). Imputation and inference in the presence of missing data. *Handbook of statist., 29A, Sample surveys: design methods and applications.* Elsevier/North Holland, Amsterdam 215-246.

[5] Särndal, C-E, Swensson B. and Wretman J. (1992). *Model assisted survey sampling*, Springer Series in Statistics. Springer-Verlag. New York, second edition.

[6] Staniswalis J.G., Lee J.J. (1998). Nonparametric Regression Analysis of Longitudinal Data, *J. Amer. Static. Assoc.* **93**, 1403-1418.