Imputation par le plus proche voisin dans des échantillons de courbes de consommation électrique incomplètes

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

[1] *Institut de Mathématiques de Bourgogne UMR CNRS 5584, Université de Bourgogne, Dijon, France.* {*herve.cardot,camelia.goga*}*@u-bourgogne.fr*
[2] *EDF R&D, Clamart, anne.de-moliner@edf.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. Afin de limiter ces phénomènes, une solution consiste à imputer ces valeurs manquantes. Dans cette communication, nous proposons donc une méthode d'imputation par le plus proche voisin permettant de compléter les portions de courbes manquantes en choisissant un donneur adapté pour chaque séquence de valeurs manquantes en fonction non seulement des caractéristiques de l'individu mais aussi du niveau de la courbe avant et après la suite de valeurs manquantes.

Mots-clés. Données fonctionnelles, 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. This problem reduces the accuracy of the estimators and may generate bias and in order to minimize these consequences, we have to impute missing values. Therefore the aim of this communication is to present a nearest neighbor imputation method which can be used to fill in the gaps in the incomplete curves, by choosing an appropriate donor for each missing values sequence depending on the characteristics of the unit and on the values of the curve before and after the missing values sequence.

Keywords. Functional data, industry, missing values, donor imputation.

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 mesurées 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,...). On pourra consulter Cardot *et al.* (2013) pour une comparaison de différentes approches en vue d'estimer la courbe de charge moyenne et de construire une bande de confiance.

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.

En particulier, la méthode des plus proches voisins est une technique d'imputation fréquemment utilisée car elle présente l'avantage d'être simple (et intuitive) et générale: il s'agit d'une approche non paramétrique qui ne suppose pas de modèle paramétrique particulier reliant la valeur manquante et les variables auxiliaires utilisées. On pourra consulter les références Haziza (2009), Shao & Wang (2008) et Beaumont & Bocci (2009) pour plus de détails et une bibliographie conséquente.

En outre, lorsqu'on impute plusieurs variables simultanément, ce qui est notre cas puisque l'on sera fréquemment confrontés à des séquences de plusieurs valeurs manquantes consécutives, un avantage majeur de cette méthode est qu'elle permet de préserver la cohérence interne de chacune des courbes complétées. Celles-ci peuvent donc être exploitées individuellement, ce qui n'est pas le cas avec l'imputation à la moyenne par exemple. En effet, même si notre but premier est d'estimer une courbe moyenne de consommation, les courbes individuelles peuvent dans certains cas être utilisées pour d'autres études.

Nous allons donc voir comment adapter au mieux ces techniques à nos problématiques de données fonctionnelles.

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 :

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

Pour cela, 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 $\pi_{kl} = \Pr(k \in s \& l \in s)$. On utilise alors l'estimateur de Horvitz-Thompson:

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

avec $w_k = \frac{1}{\pi_k}$ le poids de Horvitz-Thompson, qui ne dépend pas du temps.

3 Imputation par le plus proche voisin

Il arrive que tout ou une partie de la courbe de certains individus ne soit 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.

Pour remplacer ces valeurs manquantes, on va utiliser la méthode des plus proches voisins. Cette stratégie consiste à remplacer chaque série de valeurs manquantes par la trajectoire aux instants concernés pour un individu (ou la moyenne des trajectoires pour plusieurs individus) dont les caractéristiques sont les plus semblables possibles et dont la courbe est connue pour toute la durée du trou.

Dans notre cas, les variables auxiliaires utilisées seront la consommation aux instants juste avant et juste après le trou. [1]

En outre, afin de rendre la méthode plus efficace, on créera au préalable des classes d'imputation, qui regrouperont des clients similaires au regard des informations disponibles. On appliquera ensuite la méthode classe par classe (c'est-à-dire qu'on cherchera le ou les plus proches voisins seulement parmi les individus de la même classe que l'individu à imputer). Cela permettra de réduire les temps de calculs en limitant la recherche à un

[1] On pourrait également rajouter des variables indiquant les évolutions des consommations à ces instants encadrant le trou: $y_i(d-1) - y_i(d)$ et $y_i(f+1) - y_i(f)$ pour un trou débutant en d et se terminant en f.

nombre plus faible de clients mais aussi d'améliorer la méthode en diminuant la variance d'imputation (puisque les clients d'une même classe se ressemblent davantage).

Les propriétés théoriques permettant de justifier l'usage de la méthode des plus proches voisins pour l'imputation ont été établies par Chen & Shao (2000) sous des conditions générales sur le plan de sondage, les mécanismes de non réponse (qui doivent cependant rester non informatifs, c'est-à-dire pas liés directement à la valeur de la variable manquante) et la relation (qui doit être Lipschitzienne, dérivable suffit donc) entre la variable dépendante et la variable auxiliaire.

On notera $s_r(t)$ l'échantillon des répondants et $s_m(t)$ celui des non répondants (on remarquera au passage que ces échantillons varient avec le temps t). Formellement, l'estimateur du total après imputation s'écrira:

$$\hat{\mu}_I(t) = \sum_{k \in s_r(t)} w_k y_k(t) + \sum_{k \in s_m(t)} w_k y_{l(k)}(t) \quad (1)$$

$$= \sum_{k \in s_r(t)} W_k(t) y_k(t) \quad (2)$$

avec $l(k) \in s_r(t)$ le donneur pour $k \in s_m(t)$.

Ici nous avons fait le choix de prendre un répondant unique par séquence de valeurs manquantes: lorsque la séquence dure plusieurs instants, les valeurs imputées consécutives seront donc cohérentes entre elles. En revanche, il est possible que le voisin choisi pour deux trous différents d'une même courbe ne soit pas le même. Cela permet de s'adapter au mieux à chaque séquence de valeurs manquantes en se servant de l'information juste avant et après le trou, mais aussi d'augmenter le nombre de donneurs potentiels, puisque les donneurs ne doivent pas avoir de valeurs manquantes pendant la période à imputer.

Création des classes d'imputation

La mise en oeuvre de la méthode des plus proches voisins nécessite de constituer des classes d'imputation regroupant des clients aux courbes similaires, construites à l'aide de l'information disponible (information auxiliaire disponible dans nos bases de données et historique de la courbe). Plus précisément, dans notre contexte d'estimation de courbes de charge, ces classes seront construites sur la base des informations suivantes:

- 1. L'allure infrajournalière des consommations déduite de la courbe de charge sur l'année précédente (on estime la puissance moyenne pour chaque demi-heure que l'on divise par la puissance moyenne sur la journée, ce qui nous donne ce qu'on appelle le "profil journalier" du client, puis on classifie ces profils grâce à une méthode de classification de Kohonen. On distinguera six classes de formes, reflétant notamment la variabilité intrajournalière des consommations et leur répartition jour/nuit ainsi que la présence ou non d'un creux de mi-journée.

- 2. La sensibilité de la consommation à la température extérieure mesurée par le ratio de consommation d'hiver sur la consommation totale pour l'année précédente. Plus précisément, on créera deux classes: les clients dont la consommation d'hiver représente plus de 45% du total et les autres

- 3. La consommation totale sur l'année précédente issue de nos bases de facturation: le niveau moyen de consommations a en effet un impact prédictif très fort sur les consommations électriques à chaque instant. 5 classes de niveau de consommation seront constituées pour chaque croisement des variables précédentes.

Dans les tests réalisés, on aura donc au total soixante classes d'imputation, contenant chacune en moyenne un peu plus de 300 clients dans la population test.

Calcul de la variance

Le calcul de la variance pour l'estimateur d'un total obtenu après imputation par les plus proches voisins en présence d'une variable auxiliaire a été décrit dans Beaumont & Bocci (2009) et Shao & Wang (2008). Dans notre cadre fonctionnel une difficulté supplémentaire importante provient du fait que pour l'estimation de la trajectoire en un instant t donné, il faut considérer autant de modèles d'imputation qu'il y a de configurations de trous (date de début et date de fin) contenant une valeur manquante à cet instant. En nous limitant à un nombre faible de configurations, il est possible, en s'inspirant des travaux de Beaumont & Bissonnette (2011), d'établir une approximation de la variance de l'estimateur qui prenne en compte des combinaisons des différents modèles utilisés pour l'imputation.

4 Conclusions

Nous avons proposé ici une méthode d'estimation de courbe moyenne à partir d'un échantillon de courbes incomplètes, basée sur l'imputation par le plus proche voisin. On choisit un donneur adapté pour chaque séquence de valeurs manquantes, en fonction d'une part des caractéristiques de l'unité (dans notre contexte: le niveau de consommation, la répartition habituelle des consommations au cours de la journée et la sensibilité de la consommation aux températures extérieures) et d'autre part de la valeur de la trajectoire juste avant et juste après la séquence de valeurs manquantes. Cela permet d'exploiter l'information fournie par la partie mesurée de la courbe, les valeurs aux différents instants étant en général très corrélées.

Ce travail sera illustré sur des données réelles.

Bibliographie

[1] Beaumont, J-F. and Bissonnette, J. (2011). Variance estimation under composite imputation: The methodology behind SEVANI. *Survey Methodology*, **37**, 171-179.
[2] Beaumont, J-F. and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *The Canadian Journal of Statistics*, **37**, 400-416.
[3] Cardot, H., Dessertaine, A., Goga, C., Josserand, E. and Lardin, P. (2013). Comparison of different sample designs and construction of confidence bands to estimate the mean of functional data: An illustration on electricity consumption. *Survey Methodology*, **39**, 283-301.
[4] Chen, J. and Shao, J. (2000). Nearest Neighbor Imputation for Survey Data. *Journal of Official Statistics*, **16**, 113-131.
[5] 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.
[6] Shao, J. (2009). Nonparametric Variance Estimation for Nearest Neighbor Imputation. *Journal of Official Statistics*, **25**, 55-62.
[7] Shao, J., Wang, H. (2008). Confidence Intervals Based on Survey Data with Nearest Neighbor Imputation. *Statistica Sinica*, **18**, 281-297.