Estimation robuste de courbes moyennes de consommations électriques par sondage en population finie

Anne De Moliner [1] & Hervé Cardot [2] & Camélia 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. De nombreuses études menées à EDF R&D se basent sur l'analyse de courbes de consommations électriques moyennes pour différents groupes de clients. Ces courbes moyennes sont estimées à l'aide de panels de milliers de courbes individuelles, sélectionnées selon un plan de sondage, et mesurées au pas de temps demi-horaire. Cependant, du fait de la forte asymétrie des consommations électriques, ces échantillons contiennent fréquemment des individus atypiques, qui peuvent avoir à eux seuls un impact important sur les estimations, en particulier lorsque l'on travaille sur de petites sous-populations. Afin de limiter l'influence de ces individus atypiques, nous avons testé quatre estimateurs basés sur le concept de biais conditionnel permettant d'adapter les méthodes d'estimation robuste en sondages (Beaumont et al (2013)) au cadre des données fonctionnelles. Pour cela, on propose soit d'utiliser la notion de profondeur afin de réaliser la troncature des influences de manière cohérente entre les différents instants, soit de se ramener au cas de variables non corrélées par une Analyse en Composantes Principales Sphérique (Locantore (1999)). Ces estimateurs sont comparés entre eux et à des estimateurs non robustes sur des données réelles.

Mots-clés. Robustesse, données fonctionnelles, sondages, influence, biais conditionnel

Abstract. Many studies carried out in the French electricity company EDF are based on the mean electricity consumption curve of groups of customers. These load curves are estimated using samples of thousands of curves measured at an half hourly time step and collected according to a sampling design. Due to the skewness of the distribution of electricity consumptions, these samples often contain outliers which can have a huge impact on the estimation especially as far as small areas are concerned. Four robust estimators based on the concept of conditional bias are proposed to address this problem, by adapting robust estimation methods in finite population (Beaumont et al (2013)) to the context of functional data. More precisely we propose either to determine the functional truncation function applied to the conditional bias by using the concepts of depth, or to use a Spherical Principal Components Analysis (Locantore (1999)) to transform our data into uncorrelated real variables. These four methods are compared to each other on real datasets.

Keywords. Robustness, Functional data, sampling, influence, conditionnal bias

1 Introduction et contexte

De nombreuses études menées par les statisticiens d'EDF R&D se basent sur des courbes de consommations électriques moyennes à un pas de temps fin pour un agrégat de clients partageant des caractéristiques communes. Ces études ont pour but notamment d'appuyer les directions opérationnelles d'ERDF qui doivent équilibrer l'offre et la demande d'électricité à tout instant sur le réseau, mais aussi la direction Commerce d'EDF lorsqu'elle souhaite évaluer l'impact d'un usage ou équipement particulier sur la consommation d'électricité.

Ces courbes de charge agrégées, aussi appelées synchrones de consommation, sont estimées à l'aide de panels de quelques milliers de clients dont on mesure la consommation pour chaque demi-heure afin d'en déduire la courbe de consommation moyenne de l'ensemble de la population considérée.

Cependant, du fait de la distribution fortement asymétrique des consommations électriques, la présence d'individus atypiques dans les panels est fréquente, et ces unités peuvent avoir à eux seuls une influence forte sur les estimations. Ces problèmes rendent les estimations instables, et ce encore davantage lorsque l'on s'intéresse à des petits domaines.

Afin d'améliorer la précision de nos estimations, on se proposera donc ici d'essayer de construire des estimateurs robustes aux valeurs influentes pour nos courbes moyennes de consommation électrique, en étendant au cadre fonctionnel des méthodes préexistantes pour l'estimation robuste en population finie (Beaumont *et al* (2012)).

2 Estimation robuste sur données fonctionnelles

Notations et cadre de travail Soit une population d'intérêt U constituée de N individus. On cherche à estimer la courbe de consommation totale $\theta(t) = \sum_{i \in U} y_i(t)$ pour les différents instants $t = 1..T$ de la période d'étude. Pour cela on dispose d'un échantillon s de n individus, tiré selon un plan de sondage aléatoire. On suppose que les probabilités d'inclusion simples π_i et doubles π_{ij} de chaque individu i ou couple d'individus i, j dans l'échantillon sont connues et strictement supérieures à zéro.

Un estimateur classiquement utilisé dans ce contexte est l'estimateur de Horvitz-Thompson : $\hat{\theta}(t) = \sum_s d_i y_i(t)$ avec $d_i = \frac{1}{\pi_i}$. Lorsque l'on dispose a posteriori d'une ou plusieurs variables auxiliaires liées à notre variable d'intérêt, on peut également utiliser un estimateur par calage (Deville et Särndal (1992)).

Estimation robuste en population finie Les estimateurs ci-dessus sont sensibles à la présence de valeurs influentes, c'est-à-dire d'unités qui, pour un plan de sondage, un échantillon et un estimateur donné, ont à elles seules un impact notable sur l'estimation. Nous allons donc chercher à transformer ces estimateurs de manière à limiter l'impact de ces valeurs influentes. Pour cela, on utilisera comme mesure d'influence le biais conditionnel (Moreno-Rebollo et al (1995)). Dans notre contexte d'estimation par sondage basée

sur le plan, il représente l'espérance de l'estimateur sachant si l'individu est inclus ou non dans l'échantillon:

$$B_i = E_p(\hat{\theta} - \theta | I_i = i_i)$$

où I_i est l'indicatrice qui vaut 1 ssi l'individu i est dans l'échantillon.

Pour construire les estimateurs robustes, on utilise l'approche de Beaumont et al (2013) qui consiste à faire apparaître dans l'expression de l'estimateur les biais conditionnels des unités de l'échantillon, puis de borner ces biais afin de rendre l'estimateur plus robuste.

$$\hat{\theta}^R = [\hat{\theta} - \sum_{i \in s} \hat{B}(y_i)] + \sum_{i \in s} \psi(\hat{B}(y_i)), t = 1, ..., T$$

La question est ici de trouver une fonction ψ pertinente, qui devra notamment être bornée. Par exemple dans le cas de variables réelles, on peut utiliser la fonction de Huber

$$\psi_c(z) = sgn(z)min(z, c), c > 0$$

La constante c assurant le compromis entre biais et variance pourra être déterminée par une approche minimax, consistant à choisir la constante c qui minimise le plus grand biais conditionnel de l'estimateur robuste (en valeur absolue) sur l'échantillon. Après calcul, on obtient l'estimateur robuste : $\theta^R = \hat{\theta} - n\overline{\Delta}(c_{opt})$ avec $n\overline{\Delta}(c_{opt}) = \frac{1}{2}(\hat{B}_{min} + \hat{B}_{max})$

Nous allons maintenant présenter trois méthodes permettant d'adapter cette approche générale au cadre des données fonctionnelles.

Méthode univariée instant par instant ROBP. La solution la plus simple à mettre en oeuvre et la plus intuitive pour résoudre notre problème est d'appliquer la méthode présentée ci-dessus indépendamment sur chaque instant. Formellement, avec cette méthode, l'estimateur s'écrit:

$$\hat{\theta}^R(t) = \hat{\theta}(t) - \frac{1}{2}(\hat{B}_{min}(t) + \hat{B}_{max}(t), t = 1, ..., T)$$

Méthodes par troncature fonctionelle ROBF ACP et ROBF MBD. La méthode précédente peut potentiellement dégrader la cohérence de la courbe estimée, puisque les troncatures auront été réalisées indépendamment. Cela peut être gênant pour certaines applications qui nécessitent la préservation des liens entre instants, par exemple pour caler des modèles de prévision. Pour pallier ce problème, on va modifier la définition de la fonction ψ de manière à effectuer une troncature fonctionnelle. Pour cela, on utilise la notion de profondeur qui est une mesure du caractère atypique d'une courbe dans un jeu de données: plus elle est faible, plus la courbe est un outlier.

On utilise en particulier la Modified Band Depth de Lopez-Pintado et Romo (2009), définie comme suit :

$$MBD(Y_i) = \frac{1}{\binom{2}{n}} \frac{1}{T} \sum_{1}^{T} \sum_{j,k \ tq \ j \neq k} 1[min(Y_j(t), Y_k(t)) \leq Y_i(t) \leq max(Y_j(t), Y_j(k))]$$

Il s'agit de regarder si la courbe Y_i est fréquemment située entre les autres: plus la courbe a tendance à être incluse parmi les autres plus elle est centrale.

Une autre mesure de profondeur peut être défninie par une Analyse en Composantes Principales Sphériques (Locantore *et al* 1999)), qui est une version robuste de l'analyse en composantes principales: on détermine la distance euclidienne entre la courbe et le centre de l'espace de projection (plus une courbe sera éloignée du centre, moins elle sera profonde) : $d_i = \sqrt{\sum_{k=1}^{K} f_{i,k}^2}$ avec K le nombre de composantes principales et $f_{i,k}$ le score de l'individu i pour la composante k. On en déduit ensuite la profondeur associée: $D^{ACP}(Y_i) = -d_i$

Une fois ces profondeurs définies, on en déduit les limites de troncature de la fonction ψ de la manière suivante: on cherche la région centrale, c'est-à-dire l'enveloppe qui contient entièrement les 50% de courbes les plus profondes, et on la dilate d'un facteur α. Les portions de courbes au dessus de la limite haute de l'enveloppe (respectivement en dessous) seront remplacées par celle-ci.

Mathématiquement, soit I l'ensemble des 50% de courbes les plus profondes. Soit L l'enveloppe basse $L(t) = min_{i \in I} y_i(t)$, U l'enveloppe haute $U(t) = max_{i \in I} y_i(t)$ et m la courbe moyenne. On propose la fonction suivante:

$$\psi_\alpha(B_i)(t) = max[min(B_i(t), m(t) + \alpha(U(t) - m(t)), m(t) + \alpha(L(t) - m(t)]$$

Le choix du paramètre α permet d'assurer le compromis entre biais et variance, il a un rôle similaire au c dans l'approche ponctuelle. On le détermine donc de manière similaire, par une approche minimax (fonctionnelle): on cherche le α qui minimise le maximum sur l'échantillon des intégrales des valeurs absolues de biais conditionnels: $\alpha_{opt} = Argmin[Max_{i \in s} \sum_{t=1}^{T} |\hat{B}_i(t) - \Delta_\alpha(t)|]$ avec $\Delta_\alpha(t) = \sum_{j \in s} \psi_\alpha(\hat{B}_j(t)) - \hat{B}_j(t)$

Méthode univariée sur composantes principales ROB PACP. Une seconde façon de préserver la cohérence interne de la courbe estimée est d'effectuer un changement de base de manière à travailler sur de nouvelles variables qui seraient non corrélées, et dont les moyennes sur l'échantillon pourraient donc être estimées de manière indépendantes. Nous allons appliquer cette idée, en nous plaçant dans la base des composantes principales de l'Analyse en Composantes Principales Sphériques. Les courbes de charge s'écrivent (avec $\epsilon_i(t)$ le résidu de l'individu i à l'instant t) : $Y_i(t) = \sum_{k=1}^{K} f_{ik}\xi_k(t) + \epsilon_i(t)$ La courbe de charge moyenne s'écrit donc:
$\theta(t) = \sum_{k=1}^{K} F_k\xi_k(t) + \sum_{i \in U} \epsilon_i(t)$ avec $F_k = \frac{1}{N}(\sum_{i \in U} f_{ik})$
Pour chaque composante, on estime la moyenne des scores F_k par l'estimateur robuste ponctuel $\hat{F}_k^R$. On passe ensuite de ces scores estimés à l'estimateur final de la courbe de charge par un changement de base.

3 Application à des données réelles

On travaille sur un jeu de données issu d'une expérimentation de la Commission de Régulation de l'Energie irlandaise sur l'impact des compteurs communicants. Il contient 3994 courbes de charges de clients résidentiels, sans valeurs manquantes, pour la semaine du 18 au 24 janvier 2010, au pas demi-horaire. On construit de l'information auxiliaire à partir de ces courbes: la consommation totale sur les six mois précédents servira de variable de stratification et la consommation totale sur la période d'étude de variable de calage.

Afin d'évaluer la qualité de nos estimateurs, on tire, pour différents plans de sondage et différentes tailles d'échantillon, un grand nombre d'échantillons dans cette population, puis on estime la courbe moyenne à partir de ces échantillons par les différents estimateurs, et on compare ces estimations à la courbe moyenne réelle.

On définit différents cas tests: le sondage aléatoire simple, le sondage stratifié sur les classes de consommation avec allocation optimale basée sur la consommation semestrielle (avec 0%, 10% ou 20 % de strata jumpers simulés) et enfin le sondage aléatoire simple avec calage. Pour chacun de ces cas tests, on testera 4 tailles d'échantillons: 400, 200, 100 et 40. Dans chaque cas, on réalisera $B = 1000$ simulations.

Indicateurs de performance On note $E_{MC}[\hat{\theta}(t)] = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^b(t)$ l'espérance Monte Carlo de l'estimateur $\hat{\theta}$ pour l'instant t, avec $\hat{\theta}^b(t)$ l'estimateur de la synchrone obtenu dans la simulation b.

Pour un instant donné, on peut construire les indicateurs de performances suivants: $RB(\hat{\theta}(t)) = 100\frac{|E_{MC}[\hat{\theta}(t)]-\theta(t)|}{\theta(t)}$ et $MSE_{MC}(\hat{\theta}(t)) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^b(t)-\theta(t))^2$ avec $\theta(t)$ la vraie valeur de la synchrone pour cet instant.

On souhaite comparer les performances des estimateurs robustes avec celles de l'estimateur non robuste $\hat{\theta}^{HT}$, on utilise donc l'indicateur de MSE normalisé: $RE(\hat{\theta}(t)) = 100\frac{MSE_{MC}[\hat{\theta}(t)]}{MSE_{MC}[\hat{\theta}^{HT}(t)]}$ Afin d'estimer la performance globale, on prend la moyenne de ces estimateurs sur l'ensemble des instants t de la période de test.

Résultats:

Les simulations montrent que les méthodes robustes permettent des gains de précision globale importants, et ce d'autant plus que l'estimateur non robuste est imprécis: ainsi les gains de précision sont plus faibles pour les plans stratifiés ou les estimateurs calés mais plus forts en présence de strata jumpers ou lorsque la taille d'échantillon diminue. De plus, la meilleure méthode robuste est la troncature ponctuelle sur les composantes principales, suivie par la méthode ponctuelle, puis par les méthodes de troncature fonctionnelle.

On ne présente ici en détail que les résultats du sondage aléatoire simple.

Critère (%)	n	HT	ROB PMNX	ROB FMBD	ROB FACP	ROB PACP
			RE			
RE	400	100	95.66	98.13	96.75	93.95
RE	200	100	92.91	97.27	95.38	91.37
RE	100	100	84.63	90.74	88.49	83.19
RE	40	100	75.67	81.26	79.65	73.84
			RB			
RB	400	0.12	-2.41	-1.86	-2.01	-1.9
RB	200	-0.25	-4.3	-3.44	-3.7	-3.47
RB	100	0.57	-5.6	-5.07	-5.12	-4.38
RB	40	0.33	-9.08	-8.54	-8.36	-6.86

Table 1: Simulation 1: SAS, sans strata jumper

4 Conclusions

Trois estimateurs ont été proposés afin d'étendre les méthodes d'estimation robuste en population finie au contexte des données fonctionnelles: l'application instant par instant des méthodes robustes univariées, la troncature fonctionnelle des biais conditionnels basées sur les notions de profondeur (Modified Band Depth ou profondeur basée sur l'ACP) et la troncature univariée appliquée aux composantes principales des courbes.

Une application sur des courbes de charges réelles a montré que ces méthodes permettent des gains de précisions notables, en particulier lorsque l'estimation est la plus instable (petits échantillons, plan de sondage n'incorporant pas d'information auxiliaire ou encore présence d'individus attribués à des mauvaises strates).

Ces travaux se poursuivront par l'étude de la problématique spécifique de l'estimation de courbes de charge pour des petits domaines.

Bibliographie

[1] Beaumont, J. F., Haziza, D., Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. Biometrika, ast010.
[2] Deville, J. C., Sarndal, C. E. (1992). Calibration estimators in survey sampling. Journal of the American statistical Association, 87(418), 376-382.
[3] López-Pintado, S., Romo, J. (2009). On the concept of depth for functional data. Journal of the American Statistical Association, 104(486), 718-734.
[4] Moreno-Rebollo, J. L., Munoz-Reyes, A., Munoz-Pichardo, J. (1999). Miscellanea. Influence diagnostic in survey sampling: conditional bias. Biometrika, 86(4), 923-928.