Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Comparaison de diff®rents plans de sondage et construction de bandes de confiance pour lœstimation de la moyenne de donn®es fonctionnelles : une illustration sur la consommation ®lectrique

par Hervé Cardot, Alain Dessertaine, Camelia Goga, Étienne Josserand et Pauline Lardin

Janvier 2014

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

• Service de renseignements statistiques	1-800-263-1136
• Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
• Télécopieur	1-877-287-4369

Programme des services de dépôt

Service de renseignements	1-800-635-7943
Télécopieur	1-800-565-7757

Comment accéder à ce produit

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, ses entreprises, ses administrations et les autres établissements. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^{s}	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Techniques d'enquête, décembre 2013
Vol. 39, No. 2, pp. 313-331
Statistique Canada, N° 12-001-X au catalogue

Comparaison de différents plans de sondage et construction de bandes de confiance pour l'estimation de la moyenne de données fonctionnelles : une illustration sur la consommation électrique

Hervé Cardot, Alain Dessertaine, Camelia Goga, Étienne Josserand et Pauline Lardin[1]

Résumé

Lorsque les variables étudiées sont fonctionnelles et que les capacités de stockage sont limitées ou que les coûts de transmission sont élevés, les sondages, qui permettent de sélectionner une partie des observations de la population, sont des alternatives intéressantes aux techniques de compression du signal. Notre étude est motivée, dans ce contexte fonctionnel, par l'estimation de la courbe de charge électrique moyenne sur une période d'une semaine. Nous comparons différentes stratégies d'estimation permettant de prendre en compte une information auxiliaire telle que la consommation moyenne de la période précédente. Une première stratégie consiste à utiliser un plan de sondage aléatoire simple sans remise, puis de prendre en compte l'information auxiliaire dans l'estimateur en introduisant un modèle linéaire fonctionnel. La seconde approche consiste à incorporer l'information auxiliaire dans les plans de sondage en considérant des plans à probabilités inégales tels que les plans stratifiés et les plans πps. Nous considérons ensuite la question de la construction de bandes de confiance pour ces estimateurs de la moyenne. Lorsqu'on dispose d'estimateurs performants de leur fonction de covariance et si l'estimateur de la moyenne satisfait un théorème de la limite centrale fonctionnel, il est possible d'utiliser une technique rapide de construction de bandes de confiance qui repose sur la simulation de processus Gaussiens. Cette approche est comparée avec des techniques de bootstrap qui ont été adaptées afin de tenir compte du caractère fonctionnel des données.

Mots-clés : Bonferroni; Bootstrap; estimateur de Horvitz-Thompson; fonction de covariance; estimateur model-assisted; modèle linéaire fonctionnel; formule de Hájek.

1 Introduction

Avec le développement de procédés automatiques d'acquisition de données à des échelles de temps fines, il n'est maintenant plus inhabituel de disposer de très grandes bases de données concernant des phénomènes qui évoluent au cours du temps. Par exemple, en France, dans les années à venir, environ 30 millions de compteurs électriques vont être remplacés par des compteurs communicants. Ceux-ci seront capables de mesurer la consommation de chaque ménage et de chaque entreprise à des pas de temps potentiellement très fins (seconde ou minute) et d'envoyer ces mesures une fois par jour à un serveur central. Un autre exemple concerne les mesures d'audience des différentes chaînes de télévision. Des boitiers, reliés à internet, permettent de mesurer en temps continu si la télévision est allumée et quelle chaîne est regardée.

L'unité statistique étudiée est alors une fonction (du temps, de l'espace), ce qui nécessite d'introduire des outils d'analyse fonctionnelle. Bien que présent dès les années 1970s (Deville 1974), Dauxois et Pousse (1976), ce domaine de la statistique s'est réellement développé au cours des années 1990, avec les

1. Hervé Cardot, Université de Bourgogne, Institut de Mathématiques de Bourgogne, 9 av. Alain Savary, 21078 DIJON, FRANCE; Alain Dessertaine, LA POSTE - DIRECTION DU COURRIER - DFI – DCPES, 2 Boulevard Newton 77543 MARNE LA VALLEE CEDEX 2 and EDF, R&D, ICAME-SOAD, 1 av. du Général de Gaulle, 92141 CLAMART, France; Camelia Goga, Université de Bourgogne, Institut de Mathématiques de Bourgogne. Courriel : camelia.goga@u-bourgogne.fr; Étienne Josserand, Université de Bourgogne, Institut de Mathématiques de Bourgogne; Pauline Lardin, Université de Bourgogne, Institut de Mathématiques de Bourgogne and EDF, R&D, ICAME-SOAD.

progrès de l'informatique. Les applications concernent des domaines divers tels que la climatologie, l'économie, la télédétection, la médecine ou encore la chimie quantitative. Le lecteur pourra se reporter aux références récentes Ramsay et Silverman (2005) et Ferraty et Romain (2011) pour un panorama des différentes techniques et des exemples d'applications.

Lorsque les bases de données potentielles sont très grandes, il peut être difficile et coûteux de collecter, de sauvegarder et d'analyser l'ensemble des données. Si de plus on s'intéresse à des indicateurs simples tels que la courbe moyenne sous des contraintes d'espace mémoire ou de coût de transmission, l'emploi de techniques de sondage afin d'extraire un échantillon peut fournir une estimation précise à un coût raisonnable (Dessertaine 2008).

Les travaux combinant analyse des données fonctionnelles et théorie des sondages sont encore peu nombreux dans la littérature statistique. Cardot, Chaouch, Goga et Labruère (2010) s'intéressent à l'analyse en composantes principales en vue de réduire la dimension des données tandis que Cardot et Josserand (2011) étudient des propriétés de convergence uniforme d'estimateurs de Horvitz-Thompson de courbes moyennes. On peut également citer Chaouch et Goga (2012) qui proposent un estimateur robuste de courbes centrales.

L'objectif de ce travail est de comparer, sur un exemple réel, différentes stratégies d'échantillonnage dans un contexte fonctionnel. Ces données réelles portent sur les consommations électriques, relevées toutes les demi-heures pendant deux semaines, d'une population test de $N = 15\,069$ compteurs électriques. Le profil temporel de consommation électrique des particuliers dépend de covariables telles que les caractéristiques météorologiques (température, *etc.*) ou géographiques (altitude, latitude ou longitude). Ces variables ne sont malheureusement pas disponibles pour cette étude et nous n'utilisons qu'une seule variable comme information auxiliaire : la consommation moyenne de chaque compteur lors de la semaine précédente. Cette information peut être facilement transmise par tous les compteurs de la population.

L'extension au cadre fonctionnel des méthodes d'estimation qui prennent en compte de l'information auxiliaire n'est pas toujours directe. Cardot et Josserand (2011) proposent de stratifier la population des courbes pour améliorer l'estimation de la courbe moyenne. Chaouch et Goga (2012), qui s'intéressent à la courbe médiane, suggèrent d'utiliser un plan proportionnel à la taille avec remise ainsi que l'estimateur poststratifié. Nous proposons dans cet article de comparer plusieurs stratégies qui permettent de prendre en compte l'information auxiliaire. Une première stratégie fait intervenir l'information auxiliaire au niveau de la sélection de l'échantillon : tirage avec un plan à probabilités inégales (stratifié, πps) et estimation avec l'estimateur de Horvitz-Thompson. La deuxième stratégie fait intervenir cette information au niveau de l'estimation : tirage avec un échantillonnage aléatoire simple sans remise et estimation en utilisant un modèle de régression linéaire (Särndal, Swensson et Wretman (1992) adapté au cadre fonctionnel (Faraway 1997).

Une nouvelle question liée au caractère fonctionnel des données apparaît alors de manière naturelle : comment quantifier l'incertitude liée à l'échantillonnage ? La question, centrale pour les sondeurs, de la construction d'intervalles de confiance, n'a été que peu abordée en statistique des données fonctionnelles où il faut alors construire des bandes de confiance. En nous inspirant de techniques basées sur l'estimation de la fonction de covariance de l'estimateur (voir Faraway (1997), Cuevas, Febrero et Fraiman (2006) ou plus récemment Degras (2011)), nous proposons tout d'abord de construire des bandes de confiance par simulation de processus gaussiens. Une justification asymptotique de la validité de ces techniques est

donnée dans Cardot, Degras et Josserand (2013) lorsque les hypothèses du théorème central limite sont vérifiées et que l'on dispose d'un estimateur précis de la fonction de covariance. Une deuxième méthode de construction, qui repose sur les techniques de bootstrap, est également mise en œuvre. Il existe essentiellement trois techniques de bootstrap en population finie : le bootstrap sans remise proposé par Gross (1980), le « rescaling » bootstrap (Rao et Wu 1988) et le « mirror-match » bootstrap (Sitter 1992). Dans ce travail, nous utilisons le bootstrap sans remise qui repose sur les adaptations pour les plans stratifiés et proportionnels à la taille proposées par Chauvet (2007).

Nous introduisons dans la seconde section les notations, les estimateurs de la courbe moyenne en présence d'information auxilaire ainsi que les estimateurs de leur fonction de covariance. Les algorithmes de construction des bandes de confiance, de type bootstrap ou par simulation de processus gaussiens, sont décrits dans la section 3. La section 4 propose ensuite une comparaison des différentes stratégies, en termes de précision des estimateurs, de largeur et de couverture des bandes de confiance et de temps de calcul, de l'estimation des courbes de charge de l'opérateur français EDF (Electricité de France). Nous considérons pour cela des échantillons de taille $n = 1\,500$ dans notre population test constituée de $N = 15\,069$ courbes. Pour finir, nous présentons quelques perspectives de recherche dans la section 5.

2 Données fonctionnelles en population finie

Considérons une population finie $U = \{1, ..., N\}$ de taille N et supposons que, pour chaque élément k de la population U, nous pouvons observer la courbe déterministe $Y_k = \left(Y_k(t)\right)_{t\in[0,T]}$. L'objectif est d'estimer la courbe moyenne de la population qui est définie pour tout instant $t \in [0, T]$, par

$$\mu(t) = \frac{1}{N}\sum_{k\in U} Y_k(t).$$

Soit s un échantillon de taille fixée n, choisi aléatoirement dans U, selon un plan de sondage $p(.)$. Soient $\pi_k = \Pr(k \in s)$ et $\pi_{kl} = \Pr(k \,\&\, l \in s)$ les probabilités d'inclusion d'ordre un et deux respectivement. On suppose que $\pi_k > 0$ pour tout élément k de la population U.

La courbe moyenne μ est estimée à l'aide de l'estimateur de Horvitz-Thompson (Cardot et coll. 2010) comme suit

$$\hat{\mu}(t) = \frac{1}{N}\sum_{k\in s}\frac{Y_k(t)}{\pi_k} = \frac{1}{N}\sum_{k\in U}\frac{Y_k(t)}{\pi_k}1_{k\in s}, \quad t \in [0, T], \tag{2.1}$$

où $1_{k\in s}$ est l'indicatrice d'appartenance de l'unité k à l'échantillon s. Pour chaque instant $t \in [0, T]$, l'estimateur $\hat{\mu}(t)$ est sans biais pour $\mu(t)$, c'est à dire $E(\hat{\mu}(t)) = \mu(t)$ où l'espérance est considérée par rapport au plan de sondage.

Généralement les trajectoires $Y_k(t)$ ne sont pas observées continûment pour $t \in [0, T]$ mais uniquement sur un ensemble de D instants de mesure $0 = t_1 < t_2 < \ldots < t_D = T$. Une stratégie classique en analyse des données fonctionnelles consiste à effectuer une interpolation ou un lissage des

trajectoires discrétisées afin d'obtenir des objets qui sont réellement des fonctions (Ramsay et Silverman 2005). Cela permet également de traiter des courbes dont les instants de mesure ne sont pas identiques. Dans le cadre des sondages, l'interpolation linéaire, lorsqu'il n'y a pas d'erreur de mesure aux points discrétisés, a été étudiée par Cardot et Josserand (2011) tandis que des procédures de lissage sont proposées dans Cardot et coll. (2013). Si le nombre de points de discrétisation est suffisant et les trajectoires sont assez régulières (mais pas nécessairement dérivables), l'erreur d'approximation due au lissage ou à l'interpolation est négligeable face à l'erreur d'échantillonnage. On suppose dans la suite que les trajectoires sont observées en tout point t de l'intervalle $[0,T]$.

La fonction de covariance de type Horvitz-Thompson $\gamma(r,t) = \text{cov}\left(\hat{\mu}(r), \hat{\mu}(t)\right)$ est donnée par

$$\gamma(r,t) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{Y_k(r)}{\pi_k}\frac{Y_l(t)}{\pi_l}$$

pour tout $(r,t) \in [0,T]\times[0,T]$ et $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$. Si on suppose que les probabilités d'inclusion d'ordre deux satisfont $\pi_{kl} > 0$, un estimateur sans biais de $\gamma(r,t)$ est donné par l'estimateur sans biais de la variance de type Horvitz-Thompson,

$$\hat{\gamma}(r,t) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{Y_k(r)}{\pi_k}\frac{Y_l(t)}{\pi_l} \tag{2.2}$$

pour tout $(r,t) \in [0,T]\times[0,T]$.

2.1 Prise en compte d'information auxiliaire pour l'estimation de la trajectoire moyenne

Il est bien connu que l'utilisation d'une information auxiliaire qui explique bien la variable d'intérêt peut beaucoup améliorer la précision de l'estimateur de Horvitz-Thompson. Dans le cas des données EDF, la température extérieure ou le type de contrat pourraient sans doute être des variables auxiliaires intéressantes. Une stratification selon la position géographique permettrait également d'obtenir des estimations pour les différentes régions. Dans cette étude, nous disposons comme variable auxiliaire de la consommation électrique totale de la semaine précédente. Nous supposons que cette variable (réelle) est observée pour tous les éléments de la population.

Nous présentons dans cette section l'estimateur de Horvitz-Thompson pour la courbe moyenne ainsi qu'une estimation de la fonction de covariance de cet estimateur pour le sondage stratifié avec échantillonnage aléatoire simple sans remise (ÉASSR) dans chaque strate, noté dans la suite STRAT, et pour l'échantillonnage proportionnel à la taille sans remise que l'on note πps. Nous considérons également un estimateur de la courbe moyenne assisté par un modèle linéaire fonctionnel.

2.1.1 Le sondage stratifié avec ÉASSR dans chaque strate (STRAT)

La population U est supposée être stratifiée en un nombre fixé H de strates $U_1,\ldots,U_H$ de tailles $N_1,\ldots,N_H$. À l'intérieur de chaque strate U_h, on tire un échantillon s_h de taille n_h selon un plan ÉASSR.

Notons $\mu_h(t) = \sum_{k \in U_h} Y_k(t) / N_h$, pour $t \in [0,T]$, la courbe moyenne dans chaque strate et $\hat{\mu}_h(t) = \sum_{k \in s_h} Y_k(t) / n_h$, son estimation. L'estimateur de la courbe moyenne μ est alors défini par

$$\hat{\mu}_{\text{strat}}(t) = \frac{1}{N}\sum_{h=1}^{H} N_h \hat{\mu}_h(t) = \sum_{h=1}^{H} \frac{N_h}{N}\left(\frac{1}{n_h}\sum_{k \in s_h} Y_k(t)\right), \; t \in [0,T]. \tag{2.3}$$

L'estimateur de Horvitz-Thompson de la fonction de covariance γ est alors

$$\hat{\gamma}_{strat}(r,t) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{Y(r)Y(t),s_h} \quad r,t \in [0,T], \tag{2.4}$$

où

$$S_{Y(r)Y(t),s_h} = \frac{1}{n_h - 1}\sum_{k \in s_h}\left(Y_k(r) - \hat{\mu}_h(r)\right)\left(Y_k(t) - \hat{\mu}_h(t)\right)$$

est l'estimateur de la fonction de covariance $S_{Y(r)Y(t),U_h}$ dans la strate h. Pour $r = t \in [0,T]$, on obtient l'estimateur de la fonction de variance comme suit

$$\hat{\gamma}_{strat}(r) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S^2_{Y(r),s_h},$$

où

$$S^2_{Y(r),s_h} = \frac{1}{n_h - 1}\sum_{k \in s_h}\left(Y_k(r) - \hat{\mu}_h(r)\right)^2$$

est l'estimateur de la variance $S^2_{Y(r),U_h}$ dans la strate h. Cardot et Josserand (2011) proposent une extension, au cadre fonctionnel, de l'allocation optimale de Neyman. Les tailles n_h des échantillons s_h vérifiant

$$n_h = n\frac{N_h\sqrt{\int_0^T S^2_{Y(r),U_h}\,dr}}{\sum_{h=1}^{H} N_h\sqrt{\int_0^T S^2_{Y(r),U_h}\,dr}}, \quad h = 1,\ldots,H, \tag{2.5}$$

permettent de rendre minimale la variance intégrée, $\int_0^T \hat{\gamma}_{\text{strat}}(t)\,dt$, de l'estimateur stratifié. Cette allocation est similaire à l'allocation obtenue dans le cadre multivarié par Cochran (1977). En remplaçant la variable Y par une autre variable X connue sur toute la population et très corrélée avec la variable d'intérêt, on obtient une allocation dite x- optimale.

Remarque 2.1 *Pour $H = 1$, nous obtenons le plan aléatoire simple sans remise (ÉASSR) et la courbe moyenne $\mu(t)$ est estimée par*

$$\hat{\mu}_{éassr}(t) = \frac{1}{n}\sum_{k \in s} Y_k(t), \; t \in [0,T]. \tag{2.6}$$

L'estimateur de la fonction de covariance défini en (2.2) est alors

$$\hat{\gamma}_{éassr}(r,t) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{Y(r)Y(t),s}. \tag{2.7}$$

2.1.2 L'échantillonnage proportionnel à la taille sans remise (πps)

Les plans d'échantillonnage proportionnels à la taille avec ou sans remise sont souvent utilisés en pratique car leur efficacité est supérieure à celle de plans à probabilités égales lorsque la variable d'intérêt est plus ou moins proportionnelle à une variable auxiliaire X qui a des valeurs strictement positives.

Dans le cas des échantillons de taille fixe n tirés sans remise, il est possible de donner l'équivalent de la formule de Yates et Grundy (1953) et Sen (1953). La fonction de covariance de $\hat{\mu}$ vérifie,

$$\gamma(r,t) = -\frac{1}{2}\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U, l\neq k}(\pi_{kl} - \pi_k\pi_l)\left(\frac{Y_k(r)}{\pi_k} - \frac{Y_l(r)}{\pi_l}\right)\left(\frac{Y_k(t)}{\pi_k} - \frac{Y_l(t)}{\pi_l}\right),\ r,t \in [0,T]. \tag{2.8}$$

Supposons que les valeurs x_k de la variable X sont connues pour toutes les unités k de la population. Il est alors possible de définir les probabilités d'inclusion :

$$\pi_k = n\frac{x_k}{\sum_{k\in U} x_k}.$$

Des méthodes ont été proposées dans la littérature pour le cas $\pi_k > 1$ (Särndal et coll. 1992).

Les probabilités d'inclusion d'ordre deux sont en général très difficiles à calculer pour les plans πps et par conséquent, la formule (2.2) ne peut pas être utilisée. Il existe cependant une approximation asymptotique simple de la variance qui a été proposée par Hájek (1964) et qui ne fait intervenir que les probabilités d'inclusion d'ordre un. Cette approximation se révèle très performante lorsque la taille de l'échantillon est grande et l'entropie du plan de sondage proche de l'entropie maximale. Pour sélectionner l'échantillon s avec des probabilités d'inclusion π_k, l'algorithme du cube (Deville et Tillé 2004) équilibré sur la variable $\boldsymbol{\pi} = (\pi_k)_{k\in U}$ peut être utilisé. Deville et Tillé (2005) montrent que pour ce plan de sondage particulier la formule de Hàjek est très performante pour estimer la variance d'un total ou d'une moyenne. Cette formule d'approximation de la variance peut aussi être utilisée pour la covariance, qui est alors estimée par

$$\hat{\gamma}_{\pi\text{ps}}(r,t) = \frac{1}{N^2}\sum_{k\in s}(1-\pi_k)\left(\frac{Y_k(r)}{\pi_k} - \hat{R}(r)\right)\left(\frac{Y_k(t)}{\pi_k} - \hat{R}(t)\right),\ \ r,t \in [0,T], \tag{2.9}$$

où

$$\hat{R}(t) = \frac{\sum_{k\in s}\frac{Y_k(t)}{\pi_k}(1-\pi_k)}{\sum_{k\in s}(1-\pi_k)}.$$

Nous avons également utilisé le sondage systématique à probabilités inégales proposé par Madow (1949) en raison de sa simplicité d'utilisation. Il est malheureusement difficile d'estimer la variance pour ce type de plan et nous ne l'utiliserons donc pas pour construire les bandes de confiance.

2.2 L'estimateur assisté par un modèle (« model-assisted »)

Considérons p variables auxiliaires réelles $X_1, \ldots, X_p$ et soit x_{kj} la valeur de la variable X_j pour le $k^{\text{ème}}$ individu. Notons par $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})'$ le vecteur contenant les valeurs de p variables auxiliaires mesurées sur le $k^{\text{ème}}$ individu. On considère que la relation entre la variable d'intérêt et les variables auxiliaires est modélisée par le modèle de superpopulation suivant

$$\xi : Y_k(t) = \mathbf{x}_k'\boldsymbol{\beta}(t) + \varepsilon_{kt}, \quad t \in [0, T] \tag{2.10}$$

avec

$$E_\xi(\varepsilon_{kt}) = 0, E_\xi(\varepsilon_{kt}\varepsilon_{lt'}) = 0 \text{ pour } k \neq l \text{ et } E_\xi(\varepsilon_{kt}\varepsilon_{kt'}) = \sigma^2_{tt'} \text{ pour } k = l.$$

Ce modèle est une généralisation immédiate à plusieurs variables auxiliaires du modèle linéaire fonctionnel proposé par Faraway (1997).

L'estimation de $\boldsymbol{\beta}$ basée sur le modèle ξ et le plan de sondage $p(\cdot)$ est donnée par

$$\hat{\boldsymbol{\beta}}(t) = \left(\sum_{k\in s}\frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k}\right)^{-1}\sum_{k\in s}\frac{\mathbf{x}_k Y_k(t)}{\pi_k}, \quad t \in [0, T]. \tag{2.11}$$

Remarquons que les poids de sondage ne dépendent pas du temps $t \in [0, T]$. Soit $\hat{Y}_k(t) = \mathbf{x}_k'\hat{\boldsymbol{\beta}}(t)$ l'estimateur basé sur le plan de sondage de la prédiction sous le modèle ξ de $Y_k(t)$. Par analogie directe avec le cas univarié (Särndal et coll. 1992), nous obtenons finalement l'estimateur suivant pour la moyenne, pour $t \in [0, T]$,

$$\begin{aligned}\hat{\mu}_{MA}(t) &= \frac{1}{N}\sum_{k\in s}\hat{Y}_k(t) - \frac{1}{N}\sum_{k\in s}\frac{\left(\hat{Y}_k(t) - Y_k(t)\right)}{\pi_k} \\ &= \frac{1}{N}\sum_{k\in U}\frac{Y_k(t) - x_k'\hat{\boldsymbol{\beta}}(t)}{\pi_k} + \frac{1}{N}\left(\sum_{k\in U}\mathbf{x}_k'\right)\hat{\boldsymbol{\beta}}(t).\end{aligned} \tag{2.12}$$

Si le modèle ξ contient la variable constante 1, alors l'estimateur devient

$$\hat{\mu}_{MA}(t) = \frac{1}{N}\sum_{k\in U}\hat{Y}_k(t), \quad t \in [0, T]. \tag{2.13}$$

Pour r et t fixés, la covariance asymptotique de $\hat{\mu}_{MA}(r)$ et $\hat{\mu}_{MA}(t)$ peut être calculée selon la technique classique des résidus (Särndal et coll. 1992),

$$\gamma_{MA}(r,t) \simeq \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)\frac{\left(Y_k(r) - \tilde{Y}_k(r)\right)}{\pi_k}\frac{\left(Y_l(t) - \tilde{Y}_l(t)\right)}{\pi_l}, \tag{2.14}$$

où $\tilde{Y}_k(r) = \mathbf{x}_k'\tilde{\boldsymbol{\beta}}(t)$ est la prédiction de $Y_k(t)$ sous le modèle de superpopulation et $\tilde{\boldsymbol{\beta}}(t) = \left(\sum_U \mathbf{x}_k\mathbf{x}_k'\right)^{-1}\left(\sum_U \mathbf{x}_k Y_k(t)\right)$ est l'estimation de $\boldsymbol{\beta}$ au niveau de la population et $r,t \in [0,T]$. Cardot, Goga et Lardin (2013) montrent que ce résultat reste valable uniformément en $r,t \in [0,T]$.

Nous proposons comme estimateur de la fonction de covariance $\gamma_{MA}(r,t)$ l'estimateur de Horvitz-Thompson de la covariance asymptotique donnée par (2.14) où $\tilde{\boldsymbol{\beta}}(t)$ est remplacé par son estimateur $\hat{\boldsymbol{\beta}}(t)$ basé sur le plan de sondage,

$$\hat{\gamma}_{MA}(r,t) = \frac{1}{N^2}\sum_{k,l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\frac{\left(Y_k(r)-\hat{Y}_k(r)\right)}{\pi_k}\frac{\left(Y_l(t)-\hat{Y}_l(t)\right)}{\pi_l}, \quad r,t \in [0,T]. \tag{2.15}$$

Remarque 2.2 *Il est tout à fait possible de considérer un modèle de superpopulation ξ plus général que le modèle linéaire proposé ici. Des techniques d'estimation basées sur un lissage par des B-splines (Goga et Ruiz-Gazen 2012) peuvent alors être envisagées. Dans notre étude, la relation entre la consommation à l'instant t et la consommation moyenne de la semaine précédente est quasi linéaire (voir figure 4.1) ce qui justifie de ne pas employer ces approches nonparamétriques.*

3 Construction des bandes de confiance

Nous considérons ici des bandes de confiance pour la courbe moyenne μ qui sont de la forme

$$\mathbb{P}\left(\mu(t) \in \left[\hat{\mu}(t) \pm c_\alpha\hat{\sigma}(t)\right],\ \forall t \in [0,T]\right) = 1-\alpha, \tag{3.1}$$

où la valeur du cœfficient c_α est inconnue, et dépend du niveau de confiance $1-\alpha$ souhaité, et $\hat{\sigma}(t)$ est un estimateur de l'écart-type de $\hat{\mu}(t)$. Le calcul de c_α est basé sur le fait que sous certaines hypothèses (Cardot et coll. 2013), le processus

$$Z(t) = \left(\hat{\mu}(t)-\mu(t)\right) / \hat{\sigma}(t), \quad t \in [0,T],$$

converge vers un processus Gaussien dans l'espace des fonctions continues $\mathcal{C}([0,T])$. On a alors

$$\mathbb{P}\left(\sup_{t\in T}|Z(t)| \le c_\alpha\right) = \mathbb{P}\left(\mu(t) \in \left[\hat{\mu}(t) \pm c_\alpha\hat{\sigma}(t)\right], \forall t \in [0,T]\right) \tag{3.2}$$

et il suffit donc de déterminer c_α, le quantile d'ordre $1-\alpha$ de la variable aléatoire réelle $\sup_{t\in[0,T]}|Z(t)|$ pour construire complètement la bande de confiance. La distribution du sup de processus Gaussiens n'est connue explicitement que pour quelques cas particuliers, le mouvement brownien par exemple.

Nous proposons deux approches pour déterminer la valeur de c_α. La première repose sur une estimation directe de l'écart-type et la simulation des processus Gaussiens $Z(t)$. La seconde, qui ne nécessite pas de disposer d'estimateur de la variance, repose sur des techniques de ré-échantillonnage où à la fois l'écart-type et la valeur de c_α sont obtenus à partir des réplications bootstrap.

3.1 Construction de bandes de confiance par simulation de processus Gaussiens

Les étapes de l'algorithme sont les suivantes :

1) Tirer l'échantillon s de taille n à l'aide du plan de sondage p et calculer l'estimateur $\hat{\mu}$ ainsi que l'estimateur $\hat{\gamma}(r,t)$ de la fonction de covariance $\gamma(r,t)$, $r,t \in [0,T]$.
2) Simuler M courbes Z_m, $m = 1\ldots, M$, de même loi que Z où Z est un processus Gaussien d'espérance 0 et de fonction de covariance ρ où $\rho(r,t) = \hat{\gamma}(r,t) / \left(\hat{\gamma}(r)\hat{\gamma}(t)\right)^{1/2}$, $r,t \in [0,T]$.
3) Déterminer c_α, le quantile d'ordre $1-\alpha$ des variables, $\left(\sup_{t\in[0,T]} \left|Z_m(t)\right|\right)_{m=1,\ldots,M}$.

Cet algorithme, très rapide et facile à mettre en œuvre, a déjà été proposé, dans le cadre d'observations i.i.d. par Faraway (1997), Cuevas et coll. (2006) et Degras (2011) pour construire des bandes de confiance. On trouvera une justification asymptotique rigoureuse de cette approche dans Cardot et coll. (2013) pour l'échantillonnage dans des populations finies.

3.2 Construction des bandes de confiance par bootstrap

Dans ce travail, nous utilisons la méthode de bootstrap proposée par Gross (1980) pour l'ÉASSR et les extensions proposées par Chauvet (2007) pour les plans STRAT et πps. Elle repose sur le principe suivant : l'échantillon s est utilisé pour simuler une population fictive U^* dans laquelle nous sélectionnons plusieurs échantillons bootstrappés. La mise en œuvre de cet algorithme n'est pas immédiate lorsque le rapport $1/\pi_k$ n'est pas entier. De nombreuses variantes ont été proposées dans la litérature pour tenir compte du cas général et nous avons décidé d'adopter celle initialement proposée par Booth, Butler et Hall (1994) pour le plan d'ÉASSR.

Considérons que l'échantillon s de taille n a été sélectionné à l'aide du plan de sondage p et soit $\hat{\mu}$ l'estimateur de μ calculé à partir de s.

Algorithme général du bootstrap

1) Dupliquer chaque individu $k \in s$, $\lfloor 1/\pi_k \rfloor$ fois, où [.] désigne la partie entière. On complète la population ainsi obtenue en sélectionnant un échantillon dans s avec une probabilité d'inclusion $\alpha_k = 1/\pi_k - \lfloor 1/\pi_k \rfloor$. Soit Y_k^*, $k \in U^*$ la valeur de la variable d'intérêt sur la pseudo-population.
2) Tirer M échantillons s_m^*, $m = 1,\ldots,M$, de taille n dans la pseudo-population U^* à l'aide du plan de sondage p^* avec des probabilités d'inclusion π_k^* et calculer

$$\hat{\mu}_m^*(t) = \frac{1}{N}\sum_{k\in s_m^*} \frac{Y_k^*(t)}{\pi_k^*},\ t \in [0,T] \text{ et } m = 1,\ldots,M.$$

3) Estimer la fonction $\hat{\sigma}(t)$ par l'écart-type empirique corrigé des $\hat{\mu}_m^*(t)$, $m = 1, \ldots, M$,

$$\hat{\sigma}^2(t) = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\mu}_m^*(t) - \hat{\mu}_\bullet^*(t)\right)^2,$$

où

$$\hat{\mu}_\bullet^*(t) = \frac{1}{M}\sum_{m=1}^{M}\hat{\mu}_m^*(t) \text{ et } t \in [0,T].$$

4) Choisir c_α comme le quantile d'ordre $1-\alpha$ des variables

$$\left(\sup_{t\in[0,T]}\frac{|\hat{\mu}_m^*(t) - \hat{\mu}(t)|}{\hat{\sigma}(t)}\right)_{m=1,\ldots,M}.$$

Une technique similaire à celle utilisée lors de l'étape 4 de l'algorithme a été utilisée par Bickel et Krieger (1989) pour construire des bandes de confiance de la fonction de répartition.

Le plan d'ÉASSR utilise l'algorithme général du bootstrap pour $\pi_k^* = n/N$, et pour le plan STRAT, nous appliquons dans chaque strate U_h, pour $h = 1, \ldots, H$, l'algorithme utilisé pour le plan d'ÉASSR avec $\pi_k^* = n_h / N_h \;\; k \in U_h$. On retrouve dans ce cas, l'algorithme proposé par Booth et coll. (1994).

L'adaptation de l'algorithme du bootstrap au plan πps a été proposée par Chauvet (2007). Elle consiste à sélectionner lors de l'étape 2 de l'algorithme général, l'échantillon s^* dans U^* avec les probabilités d'inclusion

$$\pi_k^* = \frac{nx_k}{\sum_{k\in U^*} x_k}.$$

Cette modification est nécéssaire pour respecter la contrainte de taille fixe lors du rééchantillonnage. Les probabilités d'inclusion π_k^* sont également utilisées pour estimer $\hat{\mu}_m^*$ lors de l'étape 2 de l'algorithme général. La séléction d'un échantillon πps peut être réalisée en utilisant l'algorithme du cube avec la variable d'équilibrage π. Dans ces conditions, un tri aléatoire dans la population U (resp. U^*) avant le tirage de s (resp. s_m^*) est souhaitable afin d'obtenir un plan de sondage proche de l'entropie maximale (Chauvet 2007, Tillé 2011). Chauvet (2007) donne également des résultats asymptotiques concernant la convergence de l'estimateur de la variance obtenu dans le cas du bootstrap du plan πps.

Enfin, il est également possible d'adapter cet algorithme général pour estimer la fonction de variance de l'estimateur $\hat{\mu}_{MA}$. Lors de l'étape 1 de l'algorithme, on calcule également les valeurs $\mathbf{x}_k^*$ de $\mathbf{x}_k$ dans la pseudo-population U^*. En utilisant le fait que l'estimateur assisté par un modèle linéaire est une fonction nonlinéaire d'estimateurs de type Horvitz-Thompson, la valeur bootstrappée $\hat{\mu}_{MA}^*$ de $\hat{\mu}_{MA}$ sur l'échantillon s_m^* est calculée selon

$$\hat{\mu}_{MA}^*(t) = \frac{1}{N}\sum_{k\in s_m^*}\frac{Y_k^*(t) - \mathbf{x}_k^{*'}\hat{\boldsymbol{\beta}}^*(t)}{\pi_k^*} + \frac{1}{N}\left(\sum_{k\in U}\mathbf{x}_k'\right)\hat{\boldsymbol{\beta}}^*(t)$$

où $\hat{\boldsymbol{\beta}}^{*}(t) = \left(\sum_{s_m^*} \mathbf{x}_k^* \mathbf{x}_k^{*\prime}\right)^{-1} \sum_{s_m^*} \mathbf{x}_k^* Y_k^*(t)$. Comme le remarquent Canty et Davison (1999) le fait d'utiliser le total de la variable $\mathbf{x}_k$ sur la population U au lieu de la pseudo-population U^* conduit à de meilleurs résultats en particulier quand cette variable présente des valeurs extrêmes.

4 Étude de la courbe de consommation moyenne d'électricité

Nous disposons d'une population U composée de $N = 15\,069$ courbes de consommation électrique mesurées toutes les demi-heures pendant deux semaines consécutives. Nous avons $D = 336$ points de mesure pour chaque semaine et nous souhaitons estimer la courbe moyenne de consommation de la deuxième semaine. On note $\mathbf{Y}'_k = \left(Y_k(t_1), \ldots, Y_k(t_D)\right)$, la consommation d'électricité de l'individu $k \in U$ mesurée la deuxième semaine et $\mathbf{X}'_k = \left(X_k(t_1), \ldots, X_k(t_D)\right)$ sa consommation au cours de la première semaine. La consommation moyenne de chaque individu k durant la première semaine, $x_k = \sum_{d=1}^{D} X_k(t_d) / D$, qui est une information simple et peu coûteuse à transmettre, sera utilisée comme information auxiliaire. Cette variable (réelle) qui est connue pour tous les éléments k de la population est fortement liée à la courbe de consommation courante. On note sur la figure 4.1 que la consommation courante en chaque t est quasiment proportionnelle à la consommation moyenne de la semaine précédente.

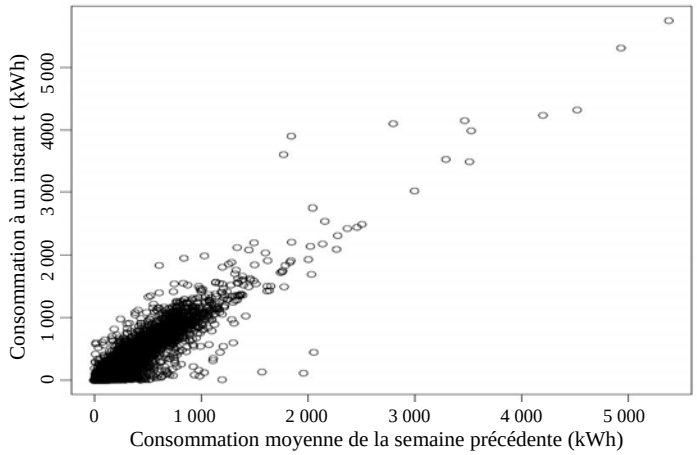


Figure 4.1 Représentation de la consommation à un instant t en fonction de la consommation moyenne de la semaine précédente

4.1 Description des stratégies utilisées

Nous considérons des échantillons de taille fixe $n = 1\,500$ obtenus selon différents plans de sondage. Les stratégies présentées sont répétées I fois afin d'évaluer et de comparer leurs performances.

1. *ÉASSR et estimateur de Horvitz-Thompson.*

La mise en œuvre de ce plan est simple, l'estimateur de Horvitz-Thompson de la courbe moyenne est donné par (2.6) et l'estimateur de sa covariance par (2.7).

2. *Sondage stratifié STRAT et estimateur de Horvitz-Thompson*.

Le plan stratifié est très efficace si les strates sont homogènes par rapport à la variable d'intérêt. Dans ce travail, nous avons utilisé l'algorithme des k -means afin de constituer les strates et nous avons considéré $H = 10$ strates. Une première stratification (STRAT 1) a été effectuée à partir de la classification des trajectoires discrétisées $\mathbf{X}'_k$ de la première semaine. Une seconde stratification, qui utilise uniquement l'information agrégée x_k a également été considérée. Elle est notée STRAT 2.

Les tailles des strates N_h obtenues en utilisant les deux stratifications ainsi que les tailles n_h optimales, selon (2.5), des échantillons à sélectionner dans chaque strate sont données dans les tableaux 4.1 et 4.2. Dans les deux cas, les strates sont numérotées en ordre croissant par rapport à la consommation moyenne de chaque strate. Plus précisement, la strate 1 correspond aux faibles consommateurs et la strate 10 est composée des 10 plus gros consommateurs d'électricité. Notons que la première stratification, qui nécessite de connaître la consommation d'électricité à chaque instant de mesure t, exige plus d'information que la deuxième stratification. La courbe moyenne est construite en utilisant (2.3) et sa covariance est estimée par (2.4).

Tableau 4.1
STRAT 1 : stratification à partir des courbes. Les strates sont construites à partir des courbes de la semaine 1. L'allocation n_h optimale est calculée à partir des courbes de la semaine 1.

h	1	2	3	4	5	6	7	8	9	10
N_h	3 866	4 769	623	2 690	664	1 251	806	328	62	10
n_h	212	345	87	242	117	179	172	101	35	10

Tableau 4.2
STRAT 2 : stratification à partir de la consommation moyenne x_k. L'allocation optimale n_h est calculée à partir de la consommation moyenne de la semaine 1.

h	1	2	3	4	5	6	7	8	9	10
N_h	3 257	4 236	3 139	1 937	1 189	731	415	125	30	10
n_h	260	293	248	204	159	133	111	56	26	10

3. *Sondage πps et estimateur de Horvitz-Thompson.*

Nous avons utilisé l'algorithme du cube proposé par Deville et Tillé (2004) et Chauvet et Tillé (2006) où les probabilités d'inclusion sont proportionnelles à x_k, $k \in U$. Afin d'avoir un plan de sondage proche de l'entropie maximale, un tri aléatoire de la population est effectué avant le tirage de l'échantillon s. La covariance de l'estimateur de la moyenne est estimée à l'aide de la formule (2.9). L'algorithme du cube est disponible sous R dans le package *sampling*, fonction *samplecube* et une macro SAS est disponible sur le site web de l'INSEE (Institut National de Statistique et des Etudes Economiques).

4. *ÉASSR et estimateur MA.*

L'estimateur $\hat{\mu}_{MA}$ assisté par le modèle ξ est construit à l'aide de l'information auxiliaire donnée par $\mathbf{x}'_k = (1, x_k)$ où x_k est la consommation moyenne de la semaine précédente. Dans ces conditions, $\hat{\mu}_{MA}$ est la somme sur toute la population U des valeurs estimées $\hat{Y}_k$ par le modèle (voir formule (2.13)). La covariance de l'estimateur de la moyenne est estimée à l'aide de la formule (2.15).

4.2 Erreur d'estimation de la courbe moyenne

L'erreur d'estimation de la courbe moyenne μ aux instants $t_1, \ldots, t_{336}$, est évaluée selon le critère

$$R_2\left(\hat{\mu}\right) = \frac{1}{336}\sum_{i=1}^{336}\left(\hat{\mu}\left(t_i\right) - \mu\left(t_i\right)\right)^2 \approx \frac{1}{T}\int_0^T \left(\hat{\mu}\left(t\right) - \mu\left(t\right)\right)^2 dt.$$

Les résultats sont présentés dans le tableau 4.3 pour $I = 10\,000$ simulations (réplications). Ils montrent clairement que, pour cette étude, la prise en compte de la consommation totale de la semaine précédente permet d'améliorer de manière importante la précision de l'estimation de la moyenne par rapport à l'échantillonnage aléatoire simple sans remise en divisant l'erreur quadratique moyenne R_2 par 5. Parmi les différentes stratégies, les plus performantes semblent être celles qui prennent en compte l'information auxiliaire via les probabilités d'inclusion (STRAT, πps et systématique proportionnel à la taille).

Tableau 4.3
Erreur quadratique R_2 d'estimation de la moyenne μ, avec $I = 10\,000$ réplications.

Stratégie	moyenne	1er quartile	médiane	3ème quartile
ÉASSR	40,53	10,82	22,16	51,09
STRAT (1)	5,78	3,68	5,08	7,07
STRAT (2)	6,49	4,03	5,48	7,88
πps	7,06	3,99	5,52	8,16
$\pi - ps$ systématique	6,73	3,85	5,20	8,07
MA	8,29	5,24	7,14	10,06

4.3 Taux de couverture et largeur des bandes de confiance

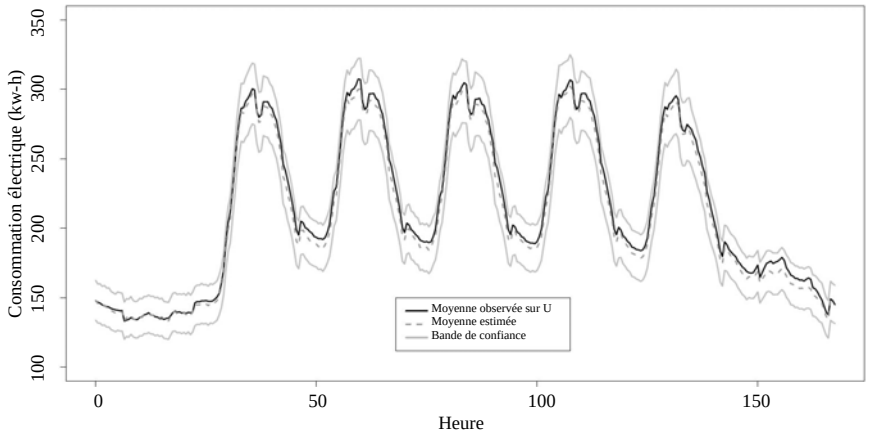
La construction des bandes de confiance de niveau $1 - \alpha$ nécessite le calcul des quantiles d'ordre $1 - \alpha$ du supremum de processus Gaussiens.

Pour ne pas privilégier une méthode de construction de bande de confiance par rapport à l'autre, nous avons appliqué les deux algorithmes sur un même échantillon s et nous avons considéré le même nombre M de processus. Ce nombre M varie d'un estimateur à l'autre en raison des temps de calculs nécessaires pour les approches de type bootstrap (voir Section 4.4).

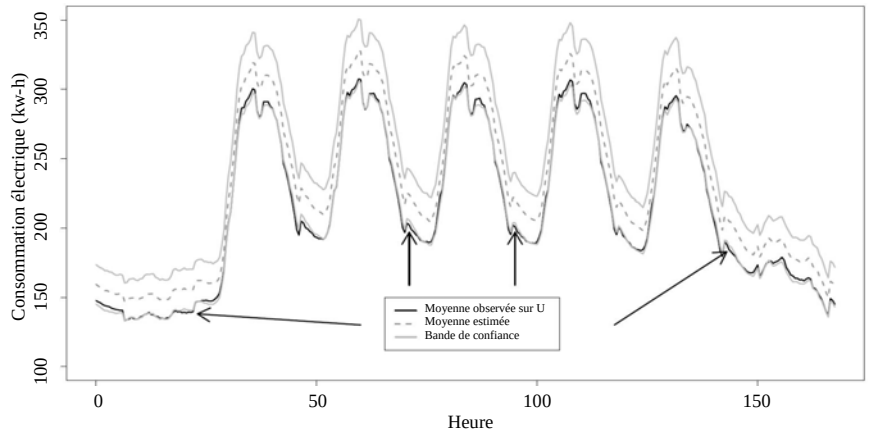
Le taux de couverture empirique est la proportion de fois, parmi les $I = 2\,000$ réplications, où la vraie courbe moyenne μ se trouve, pour tous les instants t, à l'intérieur de la bande de confiance construite à partir d'une estimation $\hat{\mu}$. Nous avons représenté sur la figure 4.2 deux exemples de bandes de confiance (courbes grises continues) construites à partir des courbes estimées (courbes grises pointillées). Sur la figure 4.2(A), nous constatons que la vraie courbe moyenne sur la population (courbe noir continue) est à l'intérieur de la bande de confiance à chaque instant. À l'opposé, sur la figure 4.2(B), nous constatons que la courbe moyenne de la population est en général surestimée et qu'il existe quelques instants (indiqués par les flèches) où la courbe observée sort de la bande de confiance. Les taux de couverture empiriques sont présentés dans le tableau 4.4.

Les deux méthodes de construction des bandes de confiance donnent des taux de couverture similaires et assez proches des taux nominaux souhaités (95 % et 99 %). Les résultats semblent cependant

légèrement moins satisfaisants pour les plans πps et pour l'approche MA pour lesquels la variance de l'estimateur est complexe et plus difficile à estimer précisément.



(A) La courbe moyenne observée est à l'intérieur de la bande de confiance.



(B) La moyenne observée est à l'extérieur de la bande de confiance aux instants indiqués par les flèches.

Figure 4.2 Exemples de bande de confiance

Tableau 4.4
Taux de couverture empirique (en %), pour $I = 2\,000$ réplications.

Méthodes	Nombre M de processus	Bootstrap		Processus Gaussien	
		$\alpha = 0,05$	$\alpha = 0,01$	$\alpha = 0,05$	$\alpha = 0,01$
ÉASSR	5 000	94,95	98,85	94,80	98,70
STRAT (1)	5 000	93,92	98,34	94,09	98,43
STRAT (2)	5 000	94,3	98,45	94	98,55
πps	1 000	94,73	98,77	93,87	98,61
MA	5 000	94,3	98,5	92,85	98,15

Un autre indicateur intéressant est la largeur moyenne de la bande de confiance,

$$\frac{1}{336}\sum_{i=1}^{336} 2c_\alpha \hat{\sigma}\left(t_i\right) \approx \frac{1}{T}\int_0^T 2c_\alpha \hat{\sigma}\left(t\right) dt$$

dont les valeurs sont présentées dans le tableau 4.5. Les deux méthodes fournissent des bandes de confiance dont les largeurs sont similaires. On note également que l'utilisation de la variable auxiliaire permet de diminuer sensiblement la largeur moyenne des bandes, celle-ci étant divisée par deux si on considère un des plans stratifiés plutôt qu'un plan d'ÉASSR.

Tableau 4.5
Largeur moyenne des bandes de confiance, pour $I = 2\,000$ réplications.

Méthodes	Nombre M de processus	Bootstrap		Processus gaussien	
		$\alpha = 0,05$	$\alpha = 0,01$	$\alpha = 0,05$	$\alpha = 0,01$
ÉASSR	5 000	35,98	43,35	35,99	43,19
STRAT (1)	5 000	16,64	18,92	16,62	18,88
STRAT (2)	5 000	17,58	19,99	17,55	19,94
πps	1 000	17,85	20,31	17,62	19,93
MA	5 000	19,88	22,65	19,75	22,44

Les figures 4.3 et 4.4 présentent les largeurs des bandes de confiance pour un niveau $\alpha = 0,05$, pour chaque instant, selon qu'elles soient ponctuelles ($c_\alpha = 1,96$), estimées par simulations de processus gaussiens ou bien obtenues en considérant l'approche basée sur l'inégalité de Bonferroni appliquée en chaque point de mesure. On a alors, dans ce dernier cas, $c_\alpha = 3,793048$, le quantile d'ordre $1 - 0,05 / (336 \times 2)$ d'une loi $N(0,1)$. Les bandes obtenues par Bonferroni sont conservatives et considèrent en quelque sorte le pire des cas en termes d'information, celui de l'indépendance des intervalles ponctuels. On peut remarquer que l'approche par simulation permet de réduire sensiblement la largeur moyenne des bandes en comparaison avec Bonferroni lorsque le plan ne permet pas de prendre en compte toute l'information temporelle des données (figure 4.3). À l'opposé, pour le plan stratifié (figure 4.4) qui permet une estimation précise de la courbe moyenne, la bande de confiance construite par simulation est proche de celle de Bonferroni, ce qui s'interprète intuitivement comme le fait que quasiment toute l'information a été capturée par le plan de sondage.

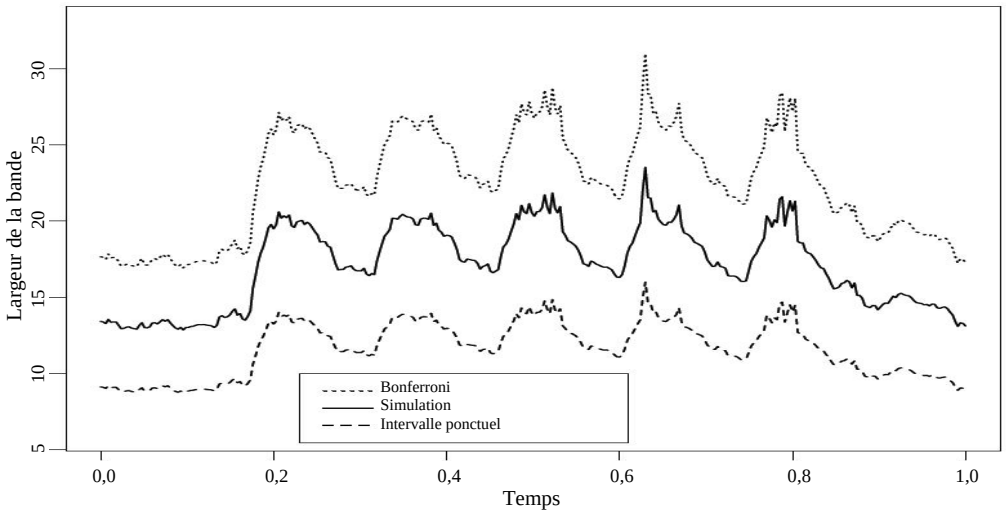


Figure 4.3 Échantillonnage aléatoire simple sans remise. Largeur des bandes de confiance ponctuelles, globales par simulations de processus et avec Bonferroni ($\alpha = 0,05$)

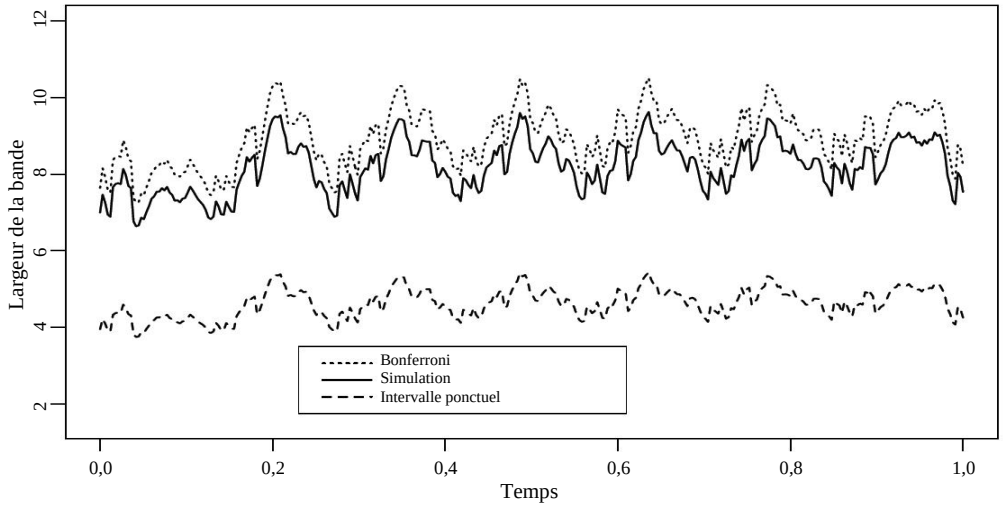


Figure 4.4 Sondage stratifié (STRAT 1). Largeur des bandes de confiance ponctuelles, globales par simulations de processus et avec Bonferroni (avec $\alpha = 0,05$)

4.4 Temps de calcul

Les temps de calcul avec la méthode par bootstrap sont largement supérieurs, de l'ordre d'un facteur de 1 à 1 000, à ceux de la méthode par simulations de processus gaussiens (voir tableau 4.6). Cette différence importante provient du fait que les méthodes de bootstrap nécessitent de répéter tout le processus d'estimation pour chaque échantillon bootstrapé : construction de la population fictive, tirage d'un nouvel échantillon, calcul de l'estimateur. On remarque également que les plans qui font intervenir de l'information auxiliaire sont moins rapides que le plan d'ÉASSR même si utilisés individuellement leur temps de calcul reste tout à fait raisonnable.

Tableau 4.6
Temps d'exécution d'une simulation en secondes pour M = 5 000 réplications. Les stratégies ÉASSR, MA et STRAT ont été programmés avec R et πps avec SAS.

Stratégie	Bootstrap	Processus gaussiens
ÉASSR	1 170,6	1,0
STRAT	1 839,5	1,4
πps	5 020,0	7,3
MA	3 156	1,4

5 Conclusion et perspectives

Nous avons, dans ce travail, mis en œuvre et comparé différentes stratégies permettant de prendre en compte de l'information auxiliaire pour l'estimation, et la construction de bandes de confiance, de la moyenne de données qui sont des courbes. Cette information peut être prise en compte au moment de l'échantillonnage en considérant des plans à probabilités inégales ou bien lors de l'estimation avec un

sondage aléatoire simple sans remise assisté par un modèle de régression à réponse fonctionnelle. Il apparaît clairement, sur notre exemple de courbes de charge d'électricité, que la connaissance des consommations totales une semaine avant, permet d'améliorer de manière importante la précision des estimateurs de la moyenne par rapport à un sondage de type ÉASSR.

Par ailleurs, dans ce contexte d'échantillons de taille importante et de données de grande dimension, il semble aussi possible de construire, pour ces différentes stratégies, des bandes de confiance qui ont des taux de couverture empiriques proches des taux souhaités. Les performances des deux approches proposées, estimation de la fonction de covariance et simulation de processus Gaussiens ou Bootstrap, semblent comparables en termes de largeur des bandes de confiance et la principale différence porte sur les temps de calcul. Le bootstrap qui semble plus général, puisqu'il ne nécessite pas de disposer d'un estimateur performant de la fonction de covariance, se révèle beaucoup plus lent en pratique.

Il y a parfois, dans ces flux de données de grande taille, des pertes d'information qui proviennent de problèmes de transmission du signal. L'opérateur observe donc au final certaines trajectoires de manière incomplète. Cette question, de non réponse partielle, peut sans doute être abordée en considérant des adaptations des techniques classiques de non réponse (Haziza 2009) au cadre fonctionnel. Une question primordiale concerne alors la construction de bons estimateurs de la fonction de covariance.

Remerciements

Nous remercions les arbitres anonymes ainsi que Guillaume Chauvet et Jean-Claude Deville pour leurs remarques fructueuses qui ont permis d'améliorer ce travail.

Bibliographie

Bickel, P., et Krieger, A. (1989). Confidence bands for a distribution function using the bootstrap. *Journal of the American Statistical Association*, 84, 95-100.

Booth, J., Butler, R. et Hall, P. (1994). Bootstrap methods for finite population. *Journal of the American Statistical Association*, 89, 1282-1289.

Canty, A.J., et Davison, A.C. (1999). Resampling-based variance estimation for labour force surveys. *The Statistician*, 48, 379-391.

Cardot, H., Chaouch, M., Goga, C. et Labruère, C. (2010). Properties of design-based functional principal components analysis. *Journal of Statistical Planning and Inference*, 140, 75-91.

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

Cardot, H., Goga, C. et Lardin, P. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic J. of Statistics*, 7, 562-596.

Cardot, H., et Josserand, E. (2011). Horvitz-thompson estimators for functional data: Asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, 98, 107-118.

Chaouch, M., et Goga, C. (2012). Using complex surveys to estimate the l1-median of a functional variable: Application to electricity load curves. *Revue Internationale de Statistique*, 80, 40-59.

Chauvet, G. (2007). Méthodes de bootstrap en population finie. Thèse de doctorat, Université de Rennes II.

Chauvet, G., et Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21, 53-61.

Cochran, W. (1977). Sampling techniques. New York: John Wiley & sons, Inc., 3ième Édition.

Cuevas, A., Febrero, M. et Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics and Data Analysis*, 51, 1063-1074.

Dauxois, J., et Pousse, A. (1976). Les analyse factorielles en calcul des probabilités et en statistique : essai d'étude synthétique. Thèse de doctorat, Université Paul Sabatier, Toulouse.

Degras, D. (2011). Simultaneous confidence bands for parametric regression with functional data. *Statistica Sinica*, 21(4), 1735-1765.

Dessertaine, A. (2008). Estimation de courbes de consommation électrique à partir des mesures synchrones. Dans *Méthodes de Sondages* (Éds., P. Guibert, D. Haziza, A. Ruiz-Gazen et Y. Tillé), Dunod, France, 353-357.

Deville, J. (1974). Méthodes statistiques et numériques de l'analyse harmonique. Ann. Insee, 15, 3-104.

Deville, J., et Tillé, Y. (2004). Efficient balanced sampling: The cube algorithm. *Biometrika*, 91, 893-912.

Deville, J., et Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Faraway, J. (1997). Regression analysis for a functional response. *Technometrics*, 39(3), 254-261.

Ferraty, F., et Romain, Y., editors (2011). *Oxford Handbook of Functional Data Analysis*. Oxford University Press.

Goga, C., et Ruiz-Gazen, A. (2013). Efficient estimation of nonlinear finite population parameters using nonparametrics, à paraitre dans le *Journal of the Royal Statistical Society*, Series B, DOI: 10.111/rssb.12024.

Gross, S. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*, 181-184.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.

Haziza, D. (2009). Imputation and inference in the presence of missing data. Dans *Sample Surveys: Theory Methods and Inference*, volume 29 de *Handbook of Statistics*, (Éds., C. Rao et D. Pfeffermann), North-Holland, 215-246.

Madow, W. (1949). *On the theory of systematic sampling*, ii. Annals of Mathematical Statistics, 19, 535-545.

Ramsay, J., et Silverman, B. (2005). *Functional data analysis*. Springer, New York, deuxième édition.

Rao, J., et Wu, C. (1988). Resampling inference with complex data. *Journal of the American Statistical Association*, 83, 231-241.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model assisted survey sampling.* Springer.

Sen, A. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5, 119-127.

Sitter, R.R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

Tillé, Y. (2011). Dix années d'échantillonnage équilibré par la méthode du cube : une évaluation. *Techniques d'enquête*, 37, 233-246.

Yates, F., et Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B, 15, 235-261.