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, Etienne Josserand, Pauline Lardin

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

H. Cardot[(a)], A. Dessertaine[(b),(c)], C. Goga[(a)], E. Josserand[(a)], P. Lardin[(a),(c)]

(a) Université de Bourgogne, Institut de Mathématiques de Bourgogne, 9 av. Alain Savary, 21078 DIJON, FRANCE

(b) LA POSTE - DIRECTION DU COURRIER - DFI - DCPES 2 Boulevard Newton 77543 MARNE LA VALLEE CEDEX 2

(c) EDF, R&D, ICAME-SOAD, 1 av. du Général de Gaulle, 92141 CLAMART, FRANCE

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 totale de la période précédente. Une première stratégie consiste à utiliser un plan de sondage simple, tel que le 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.

Abstract

When the study variable is functional and storage capacities are limited or transmission costs are high, selecting with survey techniques a small fraction of the observations is an interesting alternative to signal compression techniques. Our study is motivated, in such a functional data analysis context, by the estimation of the temporal evolution of mean electricity consumption curves over one week. We compare in this work different ways of taking auxiliary information into account. A first one consists in using simple sampling designs, such as simple random sampling without replacement, and model assisted estimators thanks to the functional linear model. A second strategy consists in considering unequal probability sampling designs such as stratified sampling or πps that can take additional information into account through their sampling weights.

Then, we address the issue of building reliable confidence bands. When consistent estimators of the covariance function of the estimators are easy to build and the mean estimator satisfies a Functional Central Limit Theorem, a fast technique based on simulations of Gaussian processes in order to approximate the distribution of their suprema can be employed. This approach is compared to bootstrap techniques which are also natural candidates for building confidence bands and that can be adapted to the finite population settings with functional data.

Mots clés : Bonferroni, Bootstrap, estimateur de Horvitz-Thompson, fonction de covariance, estimateur model-assisted, modèle linéaire fonctionnel, formule de Hájek.
Keywords : Bonferroni, Bootstrap, Horvitz-Thompson estimator,covariance function, model-assisted estimator, functional linear model, Hájek formula.

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, dans les années à venir, l'opérateur français EDF (Electricité de France) a planifié d'installer plus de 30 millions de 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. 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 1970 (Deville (1974), Dauxois and Pousse (1976)), ce domaine de la statistique s'est réellement développé au cours des années 1990, avec les 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 and Silverman (2005) et Ferraty and 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 donnés fonctionnelles et théorie des sondages sont encore peu nombreux dans la littérature statistique. Cardot et al. (2010) s'intéressent à l'analyse en composantes principales en vue de réduire la dimension des données tandis que Cardot and Josserand (2011) étudient des propriétés de convergence uniforme d'estimateurs de Horvitz-Thompson de courbes moyennes. On peut également citer Chaouch and 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 $N = 15069$ 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 considérons comme information auxiliaire la consommation de la semaine précédente. Nous comparons les performances de différents plans de sondage qui prennent en compte cette information auxiliaire afin d'améliorer la précision de l'estimateur de la courbe moyenne de consommation. Le premier type de plan fait intervenir l'information 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. Le deuxième type de plan la fait intervenir au niveau de l'estimation : tirage avec un sondage aléatoire simple sans remise et estimation à l'aide du modèle assisté basé sur la régression linéaire (Särndal et al. (1992)) adaptée au cadre fonctionnel en s'inspirant de 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 (Faraway (1997), Cuevas et al. (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 et al. (2011) 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, adaptées aux populations finies (Booth et al. (1994), Chauvet (2007)), est également mise en œuvre.

Nous introduisons dans la seconde section les notations ainsi que l'estimateur de la moyenne pour le sondage aléatoire simple sans remise. La troisième section présente les estimateurs de la courbe moyenne en présence d'information auxilaire ainsi que les estimateurs de leur fonction de covariance. Les algorithmes, de type bootstrap ou simulation de processus, de construction des bandes de confiance sont décrits dans la section 4. La section 5 propose enfin 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, sur l'estimation des courbes de charge d'EDF. Nous considérons pour cela des échantillons de taille $n = 1500$ dans notre population test constituée de $N = 15069$ courbes.

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 $Y_k = (Y_k(t))_{t\in[0,T]}$. L'objectif est d'estimer la courbe moyenne de la population qui est définie pour tout instant $t \in [0, T]$, par

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

Soit s un échantillon de taille fixée n, choisi aléatoirement dans U selon un plan de sondage $p(\cdot)$. Soient π_k =Pr($k \in s$) et π_{kl} =Pr($k\&l \in s$) les probabilités d'inclusion d'ordre un et respectivement deux. 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 al. (2010)) comme suit

$$\hat{\mu}(t) = \frac{1}{N}\sum_{k\in s}\frac{Y_k(t)}{\pi_k} = \frac{1}{N}\sum_{k\in U}\frac{Y_k(t)}{\pi_k}\mathbb{1}_{k\in s}, \quad t \in [0, T], \tag{2}$$

où $\mathbb{1}_{k\in s}$ 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.

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

$$\gamma(r, t) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{Y_k(r)}{\pi_k}\frac{Y_l(t)}{\pi_l} \tag{3}$$

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,

$$\widehat{\gamma}(r, t) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{Y_k(r)}{\pi_k}\frac{Y_l(t)}{\pi_l} \tag{4}$$

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

Remarque 2.1. *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 $0 = t_1 < t_2 < ... < t_D = T$. 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 and 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 and Josserand (2011) tandis que des procédures de lissage sont proposées dans Cardot et al. (2011).*

Estimation sans information auxiliaire : sondage aléatoire simple sans remise (SRSWOR)

On considère un sondage aléatoire simple sans remise s de taille n dans la population U de taille N. Cela revient à sélectionner sans remise n labels parmi N et à mesurer pour chaque $k \in s$, $Y_k(t)$ pour tout instant $t \in [0,T]$. Dans le cas d'un sondage aléatoire simple sans remise, $\pi_k = n/N$ et l'estimateur de Horvitz-Thompson pour la courbe moyenne μ défini dans (2) devient

$$\widehat{\mu}_{\text{srswor}}(t) = \frac{1}{n}\sum_{k\in s} Y_k(t), \quad t \in [0,T]. \tag{5}$$

L'estimateur de la fonction de covariance défini par (4) est alors

$$\widehat{\gamma}_{\text{srswor}}(r,t) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{Y(r)Y(t),s}, \quad r,t \in [0,T], \tag{6}$$

où $S_{Y(r)Y(t),s}$ est la covariance entre $Y(r)$ et $Y(t)$ calculée dans l'échantillon s,

$$\begin{aligned} S_{Y(r)Y(t),s} &= \frac{1}{n-1}\sum_{k\in s}\left(Y_k(r) - \widehat{\mu}_{\text{srswor}}(r)\right)\left(Y_k(t) - \widehat{\mu}_{\text{srswor}}(t)\right) \\ &= \frac{1}{n-1}\left(\sum_{k\in s} Y_k(r)Y_k(t) - n\widehat{\mu}_{\text{srswor}}(r)\widehat{\mu}_{\text{srswor}}(t)\right). \end{aligned}$$

Pour $r = t$, on obtient l'estimateur de la fonction de variance,

$$\widehat{\gamma}_{\text{srswor}}(t) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{Y(t),s}, \tag{7}$$

où

$$S^2_{Y(t),s} = \frac{1}{n-1}\sum_{k\in s}\left(Y_k(t) - \widehat{\mu}_{\text{srswor}}(t)\right)^2$$

est la variance corrigée dans l'échantillon s de la variable Y mesurée à l'instant t.

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

Dans cette section, nous allons entreprendre une étude détaillée de l'estimation de la trajectoire moyenne μ en prenant en compte l'information auxiliaire. 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 la pratique, la majorité des bases de sondage contiennent des variables auxiliaires. Dans le cas de 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. Nous disposons comme variable auxiliaire pour cette étude de la consommation électrique de la semaine précédente. Cette variable est également fortement liée à la courbe de consommation courante (cf. Figure 1).

L'extension au cadre fonctionnel des méthodes d'estimation qui prennent en compte de l'information auxiliaire n'est pas toujours directe. Cardot and Josserand (2011) proposent de stratifier la population des courbes pour améliorer l'estimation de la courbe moyenne. Chaouch and Goga (2012), qui s'intéressent à la courbe médiane, suggèrent également d'utiliser un plan proportionnel à la taille avec remise ainsi que l'estimateur poststratifié. Nous proposons dans cet article d'étendre les travaux précédents en considérant plusieurs stratégies qui permettent de prendre en compte l'information auxiliaire ainsi que construire des bandes de confiance pour la courbe moyenne calculées à l'aide d'une approximation asymptotique normale ou par bootstrap.

L'information auxiliaire peut être utilisée lors du plan de sondage (stratification, tirage à probabilités inégales, tirage équilibré,..) ou lors de l'estimation (estimation assistée par un modèle ou calage). Dans cette section, nous allons donner l'estimateur de Horvitz-Thompson pour la courbe moyenne ainsi qu'une estimation de la fonction de covariance pour le sondage stratifié avec SRSWOR dans chaque strate, noté dans la suite STRAT, ainsi que le sondage proportionnel à la taille sans remise que l'on note πps. Un estimateur assisté par un modèle linéaire est également construit.

3.1 Le sondage stratifié avec SRSWOR dans chaque strate (STRAT)

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

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

$$\widehat{\mu}_{\text{strat}}(t) = \frac{1}{N}\sum_{h=1}^{H} N_h \hat{\mu}_h(t) = \sum_{h=1}^{H} \frac{N_h}{N}\left(\frac{1}{n_h}\sum_{k\in s_h} Y_k(t)\right),\ t\in[0,T]. \tag{8}$$

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

$$\widehat{\gamma}_{\text{strat}}(r,t) = \frac{1}{N^2}\sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h}\ S_{Y(r)Y(t),s_h}\ \ r,t\in[0,T], \tag{9}$$

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

$$\widehat{\gamma}_{\text{strat}}(r) = \frac{1}{N^2}\sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h}\ S^2_{Y(r),s_h}, \tag{10}$$

où $S^2_{Y(r),s_h} = \frac{1}{n_h-1}\sum_{k\in s_h}(Y_k(r)-\hat{\mu}_h(r))^2$ est l'estimateur de la variance $S^2_{Y(r),U_h}$ dans la strate h.

Le plan stratifié est d'autant plus efficace que les strates sont homogènes par rapport à la variable Y. Cardot and Josserand (2011) utilisent une méthode de classification non supervisée de type k-means pour construire des strates homogènes par rapport à une variable de stratification fonctionnelle. Ils proposent également une extension, au cadre fonctionnel, de l'allocation optimale de Neyman. Les tailles n_h des échantillons s_h vérifiant

$$n_h \ = \ n\frac{N_h\sqrt{\int_0^T S^2_{Y(r),U_h}dr}}{\sum_{h=1}^{H} N_h\sqrt{\int_0^T S^2_{Y(r),U_h}dr}},\quad h=1,\ldots,H, \tag{11}$$

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

3.2 Le sondage proportionnel à la taille sans remise (πps)

Les plans de sondage 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 lorsqu'il existe une relation linéaire forte entre la variable d'intérêt et 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 and Grundy (1953) et Sen (1953), la fonction de covariance de $\hat{\mu}$ vérifie,

$$\gamma(r,t) = -\frac{1}{2}\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U, l\neq k}(\pi_{kl}-\pi_k\pi_l)\left(\frac{Y_k(r)}{\pi_k}-\frac{Y_l(r)}{\pi_l}\right)\left(\frac{Y_k(t)}{\pi_k}-\frac{Y_l(t)}{\pi_l}\right),\ r,t\in[0,T]. \tag{12}$$

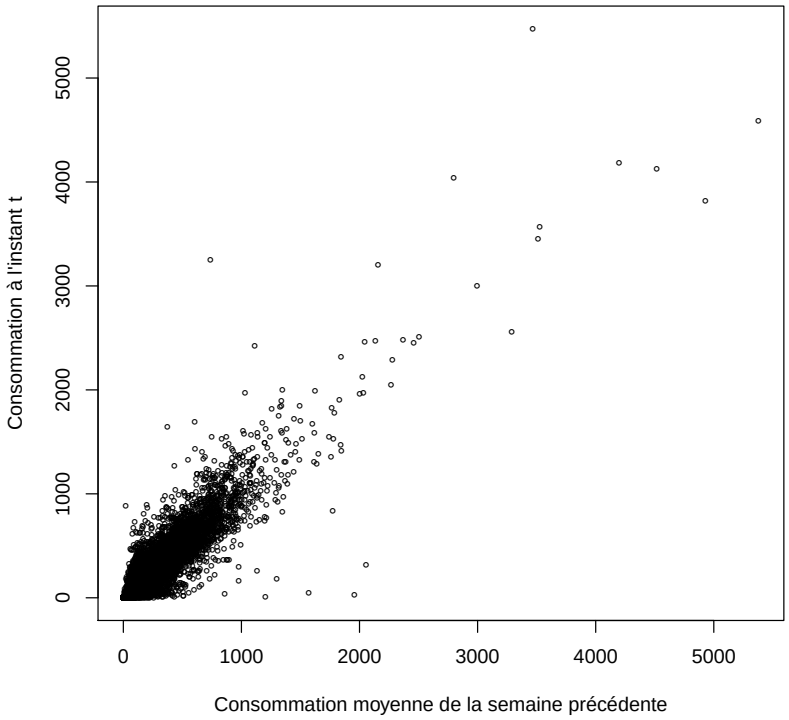


FIGURE 1 – Représentation de la consommation à un instant t en fonction de la consommation moyenne de la semaine précédente.

La variance sera donc faible, si pour tout $t \in [0,T]$ et $k \in U$, $Y_k(t)$ est proportionnelle à la probabilité d'inclusion π_k. 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 π_k par

$$\pi_k \quad = \quad n\frac{x_k}{\sum_{k\in U} x_k}. \tag{13}$$

Notons que si certaines valeurs x_k sont très élevées, cette méthode peut conduire à des $\pi_k > 1$. Dans ce cas, nous sélectionnons automatiquement ces unités et nous recalculons les probabilités d'inclusion π_k sans les individus déjà sélectionnés. Nous répétons cet algorithme jusqu'à ce que toutes les valeurs de π_k soient inférieures ou égales à 1. Nous utilisons l'estimateur (2) avec π_k définie par (13) pour estimer la courbe moyenne μ. Soit $\hat{\mu}_{\pi ps}$ l'estimateur ainsi défini.

Si les π_{kl} sont connues pour tout couple (k,l) dans s, une estimation sans biais de la fonction de covariance peut être donnée par l'équation (4). Les probabilités d'inclusion d'ordre deux sont en général très difficiles à calculer pour les plans πps. Il existe cependant une approximation asymptotique simple, en ne faisant intervenir que les probabilités d'inclusion d'ordre un, de la variance qui a été proposée par Hájek (1964). 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 d'ordre un π_k, nous avons utilisé l'algorithme du cube (Deville and Tillé (2004)) équilibré sur la variable $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$. Deville and Tillé (2005) montrent que pour ce plan de sondage particulier la formule d'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

$$\widehat{\gamma}_{\pi\mathrm{ps}}(r,t) = \frac{1}{N^2}\sum_{k\in s}(1-\pi_k)\left(\frac{Y_k(r)}{\pi_k} - \widehat{R}(r)\right)\left(\frac{Y_k(t)}{\pi_k} - \widehat{R}(t)\right), \quad r,t \in [0,T], \tag{14}$$

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

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.

3.3 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-ème individu. Notons par $\mathbf{x}_k = (x_{k1}, ..., x_{kp})'$ le vecteur contenant les valeurs de p variables auxiliaires mesurées sur le k-ème individu. On considère que la relation entre la variable d'intérêt et les variables auxiliaires est modélisée par le modèle de superpopulation suivant

$$\xi: \quad Y_k(t) \;=\; \mathbf{x}_k'\boldsymbol{\beta}(t) + \epsilon_{kt}, \quad t \in [0,T] \tag{15}$$

où $\boldsymbol{\beta}(t)$ est le vecteur des cœfficients de régression, $E_\xi(\epsilon_k) = 0$ et $\mathrm{cov}_\xi(\epsilon_k) = \boldsymbol{\Sigma}$. Ce modèle est une généralisation immédiate à plusieurs variables auxiliaires du modèle linéaire fonctionnel proposé par Faraway (1997).

Les poids de sondage $1/\pi_k$ et les couples $(\mathbf{x}_k, Y_k)$, $k \in s$, sont pris en compte pour calculer une estimation $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ comme suit :

$$\hat{\boldsymbol{\beta}}(t) = \left(\sum_{k\in s}\frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k}\right)^{-1}\sum_{k\in s}\frac{\mathbf{x}_k Y_k(t)}{\pi_k}, \quad t \in [0,T]. \tag{16}$$

Remarquons que les poids de sondage ne dépendent pas du temps $t \in [0,T]$. Par analogie directe avec le cas univarié (Särndal et al. (1992)) nous obtenons finalement l'estimateur suivant pour la moyenne,

$$\hat{\mu}_{MA}(t) \;=\; \frac{1}{N}\sum_{k\in U}\hat{Y}_k(t) - \frac{1}{N}\sum_{k\in s}\frac{(\hat{Y}_k(t) - Y_k(t))}{\pi_k}, \quad t \in [0,T], \tag{17}$$

où $\hat{Y}_k(t) = \mathbf{x}_k'\hat{\boldsymbol{\beta}}(t)$. Si le modèle ξ contient la variable constante 1, alors l'estimateur devient

$$\hat{\mu}_{MA}(t) = \frac{1}{N}\sum_{k\in U}\hat{Y}_k(t), \quad t \in [0,T]. \tag{18}$$

L'estimateur $\hat{\mu}_{MA}(t)$ est une fonction non linéaire de totaux et par conséquent, la fonction de covariance γ de $\hat{\mu}_{MA}(t)$ ne peut pas être calculée simplement à partir de la formule (4). Il est cependant possible d'obtenir des approximations asymptotiques de la covariance en utilisant des techniques de linéarisation. Des travaux sont en cours pour valider rigoureusement l'approximation suivante

$$\begin{aligned}\gamma_{\mathrm{MA}}(r,t) &\simeq \frac{1}{N^2}\mathrm{cov}\left(\sum_{k\in s}\frac{Y_k(r)-\tilde{Y}_k(r)}{\pi_k}, \sum_{k\in s}\frac{Y_k(t)-\tilde{Y}_k(t)}{\pi_k}\right)\\ &= \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}(\pi_{kl}-\pi_k\pi_l)\frac{\left(Y_k(r)-\tilde{Y}_k(r)\right)}{\pi_k}\frac{\left(Y_l(t)-\tilde{Y}_l(t)\right)}{\pi_l},\end{aligned} \tag{19}$$

où $\tilde{Y}_k(r) = \mathbf{x}_k'\tilde{\boldsymbol{\beta}}(t)$ est la prédiction de $Y_k(t)$ sous le modèle de superpopulation, $\tilde{\boldsymbol{\beta}}(t) = \left(\sum_U \mathbf{x}_k\mathbf{x}_k'\right)^{-1}\left(\sum_U \mathbf{x}_k Y_k(t)\right)$ est l'estimation de $\boldsymbol{\beta}$ au niveau de la population et $r,t \in [0,T]$. Nous proposons comme estimateur de la fonction de covariance $\gamma_{\mathrm{MA}}(r,t)$ l'estimateur de Horvitz-Thompson de la variance asymptotique donnée par (19) où $\tilde{\boldsymbol{\beta}}(t)$ est remplacé par son estimateur $\hat{\boldsymbol{\beta}}(t)$ basé sur le plan de sondage,

$$\widehat{\gamma}_{\mathrm{MA}}(r,t) = \frac{1}{N^2}\sum_{k,l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\frac{\left(Y_k(r)-\widehat{Y}_k(r)\right)}{\pi_k}\frac{\left(Y_l(t)-\widehat{Y}_l(t)\right)}{\pi_l}, \quad r,t\in[0,T]. \tag{20}$$

Remarque 3.1. *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 and 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 (cf. Figure 1) ce qui justifie de ne pas employer ces approches nonparamétriques.*

4 Construction des bandes de confiance

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

$$\mathbb{P}\left(\mu \in \{[\widehat{\mu}(t) \pm c_\alpha\,\widehat{\sigma}(t)]\,,\ \forall t\in[0,T]\}\right) = 1-\alpha, \tag{21}$$

où la valeur du cœfficient c_α est inconnue, et dépend du niveau de confiance $1-\alpha$ souhaité, et $\widehat{\sigma}(t)$ est un estimateur de l'écart-type de la variance $\gamma(t,t)$. Le calcul de c_α est basé sur le fait que sous certaines hypothèses (Cardot et al. (2011)), le processus

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

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

$$\mathbb{P}\left(\sup_{t\in T}|Z(t)|\leq c_\alpha\right) \quad = \quad \mathbb{P}\left(\mu\in\{[\widehat{\mu}(t)\pm c_\alpha\,\widehat{\sigma}(t)]\,,\,\forall t\in[0,T]\}\right) \tag{22}$$

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.

4.1 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 $\widehat{\gamma}(r,t)$ 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 $\rho(r,t)=\widehat{\gamma}(r,t)/\left(\widehat{\gamma}(r)\;\widehat{\gamma}(t)\right)^{1/2}$, $r,t\in[0,T]$.
3. Déterminer c_α, le quantile d'ordre $1-\alpha$ des variables, $\left(\sup_{t\in[0,T]}|Z_m(t)|\right)_{m=1,\ldots,M}$.

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 al. (2006) et Degras (2011) pour construire des bandes de confiance. On trouvera une justification asymptotique rigoureuse de cette approche dans Cardot et al. (2011).

4.2 Construction des bandes de confiance par bootstrap

Il existe essentiellement trois techniques de bootstrap en population finie : le bootstrap sans remise proposé par Gross (1980), le "rescaling" bootstrap (Rao and Wu (1988)) et le "mirror" bootstrap (Sitter (1992)). Dans ce travail, nous utilisons le bootstrap sans remise et le "mirror" bootstrap.

L'échantillon s est utilisé pour simuler une population fictive U^* dans laquelle nous sélectionnons plusieurs échantillons bootstrappés. Pour simuler la population fictive U^*, nous considérons la méthode proposée par Gross (1980) pour le sondage SRSWOR et les extensions proposées par Chauvet (2007) pour les plans STRAT et πps. La littérature sur le bootstrap en population finie pour des estimateurs assistés par un modèle est très réduite et

à notre connaissance, seul le travail de Helmers and Wegkamp (1998) propose un algorithme de bootstrap dans le cas de l'estimateur assisté par un modèle, $\hat{\mu}_{MA}$. Leur algorithme de bootstrap est développé pour un modèle linéaire hétéroscédastique sans intercept que nous adaptons au cas d'une réponse fonctionnelle. Contrairement aux algorithmes présentés pour les plans SRSWOR, STRAT et πps, l'algorithme de Helmers and Wegkamp ne nécessite pas de dupliquer la population U, il utilise le "mirror" bootstrap.

Algorithme général du bootstrap

1. Tirer un échantillon s de taille n à l'aide du plan de sondage p et calculer l'estimateur $\hat{\mu}$.
2. Dupliquer chaque individu $k \in s$, $1/\pi_k$ fois pour construire une population fictive U^*.
3. Tirer M échantillons s_m^*, $m = 1, \ldots, M$, de taille n dans la population fictive U^* à l'aide du plan de sondage p et calculer $\hat{\mu}_m^*(t) = \frac{1}{N} \sum_{k \in s_m^*} \frac{Y_k(t)}{\pi_k}$, $t \in [0, T]$ et $m = 1, ..., M$.
4. Estimer la fonction $\hat{\sigma}(t)$ par l'écart type empirique corrigé des $\hat{\mu}_m^*(t)$, $m = 1, ..., M$,

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

   où $\hat{\mu}_\bullet^*(t) = \frac{1}{M} \sum_{m=1}^{M} \hat{\mu}_m^*(t)$ et $t \in [0, T]$.
5. Choisir c_α comme le quantile d'ordre $1-\alpha$ des variables $\left(\sup_{t \in [0,T]} \frac{|\hat{\mu}_m^*(t) - \hat{\mu}(t)|}{\hat{\sigma}(t)} \right)_{m=1,...,M}$.

Une technique similaire a été utilisée par Bickel and Krieger (1989) pour construire des bandes de confiance de la fonction de répartition.

La mise en œuvre de la deuxième étape de cet algorithme peut poser quelques problèmes en pratique. En effet, $1/\pi_k$ est rarement un nombre entier et il faut donc adapter l'algorithme bootstrap pour obtenir une population fictive U^* de la même taille que la population U. De nombreuses variantes ont été proposées dans la littérature pour tenir compte du cas général où $1/\pi_k$ n'est pas un entier. Nous avons décidé d'adapter celle proposée par Booth et al. (1994).

Algorithme du bootstrap adapté au plan SRSWOR

1. Tirer un échantillon s de taille n par sondage aléatoire simple sans remise et calculer l'estimateur $\hat{\mu}_{\text{srswor}}$.
2. Dupliquer chaque individu $k \in s$, $[1/\pi_k]$ fois, où $[.]$ désigne la partie entière. On complète la population ainsi obtenue en sélectionnant un échantillon aléatoire simple sans remise dans s de taille $N - n[N/n]$. Nous obtenons ainsi une population fictive U^* de taille N.
3. Tirer l'échantillon s^* de taille n dans la population fictive U^* par sondage aléatoire simple sans remise et calculer $\hat{\mu}^*(t) = \frac{1}{n} \sum_{k \in s^*} Y_k(t)$, $t \in [0, T]$.

4. Répéter les étapes 2 et 3 M fois afin d'obtenir $\hat{\mu}^*_m(t)$, $t \in [0,T]$ et $m = 1, ..., M$.
5. Répéter les étapes 4 et 5 de l'algorithme général du bootstrap.

Algorithme du bootstrap adapté au plan STRAT

1. Tirer un échantillon s de taille n à l'aide d'un sondage aléatoire simple sans remise de taille n_h dans chaque strate U_h, $h = 1, ..., H$, et calculer l'estimateur $\hat{\mu}_{\text{strat}}$.
2. Pour $h = 1, ..., H$, dupliquer chaque individu $k \in s_h$, $[N_h/n_h]$ fois, où $[.]$ désigne la partie entière. On complète les unités ainsi obtenues en sélectionnant un échantillon de taille $N_h - n_h[N_h/n_h]$ dans s_h par sondage aléatoire simple sans remise. Nous obtenons ainsi la population fictive de la strate U^*_h de taille N_h. La population fictive est $U^* = \bigcup_{h=1}^{H} U^*_h$ de taille N.
3. Tirer l'échantillon s^* de taille n dans la population fictive U^* selon un plan STRAT et calculer $\hat{\mu}^*(t) = \frac{1}{N}\sum_{h=1}^{H}\sum_{k\in s^*_h}\frac{N_h}{n_h}Y_k(t)$, $t \in [0,T]$.
4. Répéter les étapes 2 et 3 M fois afin d'obtenir $\hat{\mu}^*_m(t)$, $t \in [0,T]$ et $m = 1, ..., M$.
5. Répéter les étapes 4 et 5 de l'algorithme général du bootstrap.

Algorithme du bootstrap adapté au plan πps

1. Tirer un échantillon s de taille n selon le plan de sondage proportionnel à la taille avec les probabilités d'inclusion π_k proportionnelles à x_k données par (13) et calculer l'estimateur $\hat{\mu}_{\pi\text{ps}}$.
2. Dupliquer chaque individu $k \in s$ $[1/\pi_k]$ fois, où $[.]$ désigne la partie entière. On complète les unités ainsi obtenues en sélectionnant un échantillon à l'aide du plan de sondage p et en prenant comme probabilité d'inclusion $\alpha_k = 1/\pi_k - [1/\pi_k]$, dans s. On obtient ainsi la population fictive U^* de taille N^*.
3. Tirer l'échantillon s^* de taille n dans U^* à l'aide du plan de sondage proportionnel à la taille avec les probabilités d'inclusion

$$\pi^*_k = n\frac{x_k}{\sum_{k\in U^*} x_k}$$

et calculer $\hat{\mu}^*(t) = \frac{1}{N}\sum_{k\in s^*}\frac{Y_k(t)}{\pi^*_k}$, $t \in [0,T]$.
4. Répéter les étapes 2 et 3 M fois afin d'obtenir $\hat{\mu}^*_m(t)$, $t \in [0,T]$ et $m = 1, ..., M$.
5. Répéter les étapes 4 et 5 de l'algorithme général du bootstrap.

Afin de respecter la contrainte de taille fixe lors du rééchantillonnage, nous avons bootstrapé le processus de calcul des probabilités d'inclusion. Ainsi, pour tirer l'échantillon s^* dans U^* nous utilisons les probabilités d'inclusion π^*_k. L'algorithme bootstrap que nous venons de proposer est d'autant plus efficace que le plan de sondage utilisé est proche d'un plan à entropie maximale (Chauvet (2007), Tillé (2011)).

Comme nous l'avons déjà remarqué dans la section 3.2, l'algorithme du cube avec la variable d'équilibrage $\boldsymbol{\pi}$ peut être utilisé pour tirer des échantillons selon un plan πps. L'utilisation de cet algorithme nécessite de trier aléatoirement la population U (resp. U^*) avant d'effectuer le tirage de l'échantillon s (resp. s_m^*). Ce tri est nécessaire pour obtenir un plan de sondage proche de l'entropie maximale.

Algorithme du bootstrap adapté au plan SRSWOR avec l'estimateur assisté par un modèle

On considère le modèle linéaire de superpopulation ξ présenté dans la section 3.3. Pour construire la bande de confiance par bootstrap, nous avons adapté le premier algorithme de Helmers and Wegkamp (1998) à notre cas. Il s'agît d'un "wild" bootstrap à deux degrés qui repose sur le "wild" bootstrap (Mammen (1993)) proposé dans le cas des modèles linéaires hétéroscédastiques de variance inconnue et un "mirror" bootstrap à deux degrés (Sitter (1992)).

1. Tirer un échantillon s de taille n selon un plan SRSWOR et calculer l'estimateur du cœfficient de régression $\widehat{\boldsymbol{\beta}}(t) = (\sum_s \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_s \mathbf{x}_k Y_k(t)$, ainsi que les résidus estimés $\hat{\epsilon}_k(t) = Y_k(t) - \mathbf{x}_k' \widehat{\boldsymbol{\beta}}(t)$, $t \in [0,T]$ et $k \in s$. La moyenne $\mu(t)$ est estimée par $\hat{\mu}_{\mathrm{MA}}(t)$ à partir de (17).
2. Simuler n variables aléatoires indépendantes et identiquement distribuées $Z_1, ..., Z_n$ de moyenne zéro et de variance 1 et calculer, pour chaque individu $k \in s$, les valeurs bootstrappées de Y_k,

$$Y_k^*(t) = \mathbf{x}_k' \widehat{\boldsymbol{\beta}}(t) + Z_k \widehat{\epsilon}_k(t), \quad t \in [0,T].$$

3. Poser $n' = \min\{([n^2/N]+1), n\}$ et $i = [n/n']$. L'échantillon bootstrap s^* est obtenu de la manière suivante
   (a) Tirer par sondage aléatoire simple sans remise dans s un échantillon s_1^* de taille n'.
   (b) Répéter i fois l'étape (a) de façon indépendante afin de constituer l'échantillon $s^* = s_1^* \cup ... \cup s_i^*$ de taille $n^* = n'i$.
4. Calculer $\hat{\boldsymbol{\beta}}^*(t) = (\sum_{s^*} \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_{s^*} \mathbf{x}_k Y_k^*(t)$ et construire l'estimateur "model-assisted" dans l'échantillon s^*,

$$\hat{\mu}_{MA}^*(t) = \frac{1}{N} \sum_{k \in U} \hat{Y}_k^*(t) - \frac{1}{N} \sum_{k \in s^*} \frac{(\hat{Y}_k^*(t) - Y_k(t))}{\pi_k}, \qquad t \in [0,T]$$

   où $\hat{Y}_k^*(t) = \mathbf{x}_k' \hat{\boldsymbol{\beta}}^*(t)$ est la valeur bootstrappée de la prédiction $\hat{Y}_k(t)$.
5. Répéter M fois les étapes 2, 3 et 4 afin d'obtenir $\hat{\mu}_{MA,m}^*(t)$, $m = 1, ..., M$.
6. Répéter les étapes 4 et 5 de l'algorithme général du bootstrap.

Pour simuler les variables Z_i lors de l'étape 2, nous avons utilisé la stratégie proposée par Mammen (1993) : $Z = (\delta_1 + N_1/\sqrt{2})(\delta_2 + N_2/\sqrt{2}) - \delta_1\delta_2$ où N_1 et N_2 sont deux variables aléatoires normales centrées réduites indépendantes, $\delta_1 = (3/4 + \sqrt{17}/12)^{1/2}$ et $\delta_2 = (3/4 - \sqrt{17}/12)^{1/2}$. Il faut remarquer également que lors de l'étape 2 de l'algorithme, la covariable $\mathbf{x}_k$ n'est pas répliquée car ce type de bootstrap essaie d'approcher la répartition des erreurs ε_k du modèle ξ conditionnellement à $\mathbf{x}_k$. L'étape 2 peut s'interpréter comme un tirage indépendant de n termes d'erreur répliqués $\varepsilon_k^*(t) = Z_k\widehat{\epsilon}_k(t)$ (Efron and Tibshirani (1993)) suivi du calcul des valeurs répliquées de Y_k par

$$Y_k^*(t) = \mathbf{x}_k'\widehat{\boldsymbol{\beta}}(t) + \varepsilon_k^*(t).$$

5 Etude de la courbe de consommation moyenne d'électricité

Nous disposons d'une population U composée de $N = 15069$ 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 $\mathbf{Y}_k' = (Y_k(t_1), \ldots, Y_k(t_D))$, la consommation d'électricité de l'individu $k \in U$ mesurée la deuxième semaine et $X_k' = (X_k(t_1), \ldots, X_k(t_D))$ sa consommation au cours de la première semaine. La consommation moyenne de chaque individu k, $x_k = \sum_{d=1}^{D} X_k(t_d)/D$, qui est une information simple et peu coûteuse à transmettre, sera utilisée comme information auxiliaire.

5.1 Description des stratégies utilisées

Nous considérons des échantillons de taille fixe $n = 1500$ selon différents plans de sondage.

1. *Sondage SRSWOR 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 (5) et l'estimateur de sa variance par (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 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 (11), des échantillons à sélectionner dans chaque strate sont données dans les tableaux 1 et 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. Il faut remarquer aussi que la première stratification exige plus d'information que la deuxième stratification car dans le premier cas, il faut connaître la consommation d'électricité à chaque instant de mesure t. La courbe moyenne est construite en utilisant (8) et sa variance est estimée par (10).

3. *Sondage πps et estimateur de Horvitz-Thompson.* Nous avons utilisé l'algorithme du cube proposé par Deville and Tillé (2004) et Chauvet and 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 (14). L'algorithme du cube est disponible sous  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. *Sondage SRSWOR 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 prédites $\hat{Y}_k$ par le modèle (cf. formule (18)).La covariance de l'estimateur de la moyenne est estimée à l'aide de la formule (20).

h	1	2	3	4	5	6	7	8	9	10
N_h	3866	4769	623	2690	664	1251	806	328	62	10
n_h	212	345	87	242	117	179	172	101	35	10

TABLE 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	3257	4236	3139	1937	1189	731	415	125	30	10
n_h	260	293	248	204	159	133	111	56	26	10

TABLE 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.

Ces stratégies sont répétées I fois afin d'évaluer et de comparer les performances des différentes approches envisagées.

5.2 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 les deux critères

$$R_1(\hat{\mu}) = \frac{1}{336} \sum_{i=1}^{336} |\hat{\mu}(t_i) - \mu(t_i)| \;\approx\; \frac{1}{T} \int_0^T |\hat{\mu}(t) - \mu(t)| dt$$

et

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

Les résultats sont présentés dans les Tables 3 et 4 pour $I = 10000$ 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 au sondage aléatoire simple sans remise en divisant l'erreur moyenne absolue R_1 par 5/2. Parmi les différentes stratégies, les plus performantes semblent être celles qui prennent en compte l'information auxiliaire via les probabilités d'inclusion (STRAT, π-ps et systématique proportionnel à la taille).

Stratégie	moyenne	1^{er} quartile	médiane	3^{eme} quartile
SRSWOR	5.00	2.70	4.05	6.48
STRAT (1)	1.91	1.55	1.83	2.19
STRAT (2)	2.01	1.62	1.90	2.31
πps	2.04	1.60	1.90	2.33
π-ps systématique	1.98	1.56	1.83	2.30
MA	2.29	1.85	2.17	2.61

TABLE 3 – Erreur R_1 d'estimation de la moyenne μ, avec $I = 10000$ réplications.

Stratégie	moyenne	1^{er} quartile	médiane	3^{eme} quartile
SRSWOR	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
π-ps systématique	6.73	3.85	5.20	8.07
MA	8.29	5.24	7.14	10.06

TABLE 4 – Erreur quadratique R_2 d'estimation de la moyenne μ, avec $I = 10000$ réplications.

5.3 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.

Pour ne pas privilégier une méthode de construction de bande de confiance par rapport à l'autre, nous avons appliqué les 2 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 5.4).

Le taux de couverture empirique est la proportion de fois, parmi les $I = 2000$ 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 2 deux exemples de bandes de confiance (courbes grises continues) construites à partir des courbes estimées (courbes grises pointillées). Sur la figure 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. A l'opposé, sur la figure 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 la Table 5.

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 lesquelles la variance de l'estimateur est complexe et plus difficile à estimer précisément.

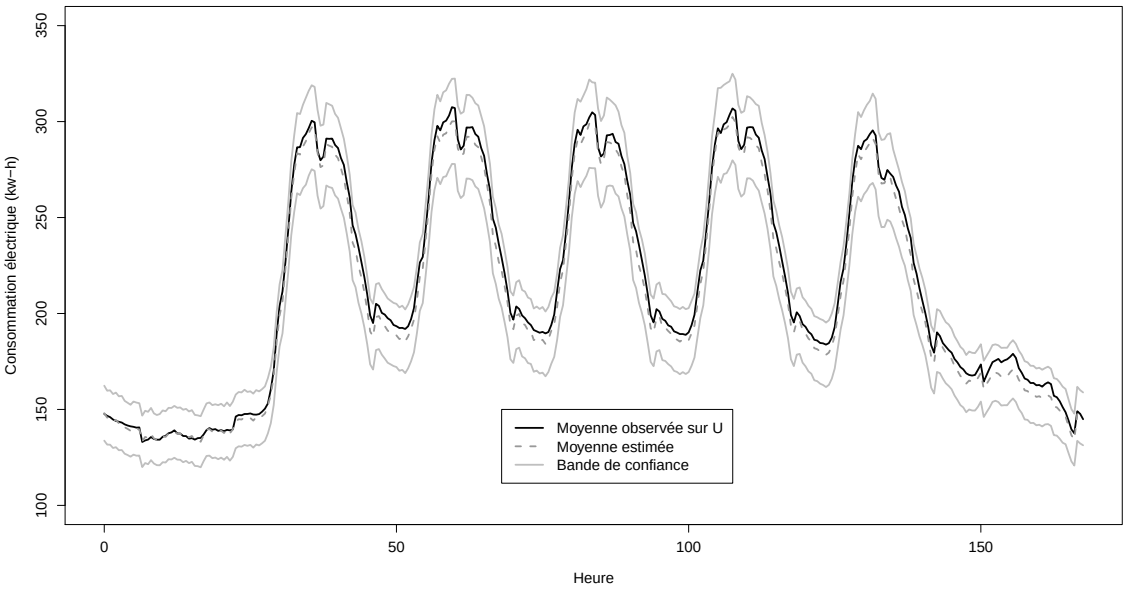
Méthodes	Nombre M de processus	Bootstrap		Processus Gaussien	
		α=0.05	α=0.01	α=0.05	α=0.01
SRSWOR	5000	94.95	98.85	94.80	98.70
STRAT (1)	5000	93.92	98.34	94.09	98.43
STRAT (2)	5000	94.3	98.45	94	98.55
πps	1000	94.73	98.77	93.87	98.61
MA	5000	94.4	99.05	93.15	98.70

TABLE 5 – Taux de couverture empirique (en %), pour I=2000 réplications.

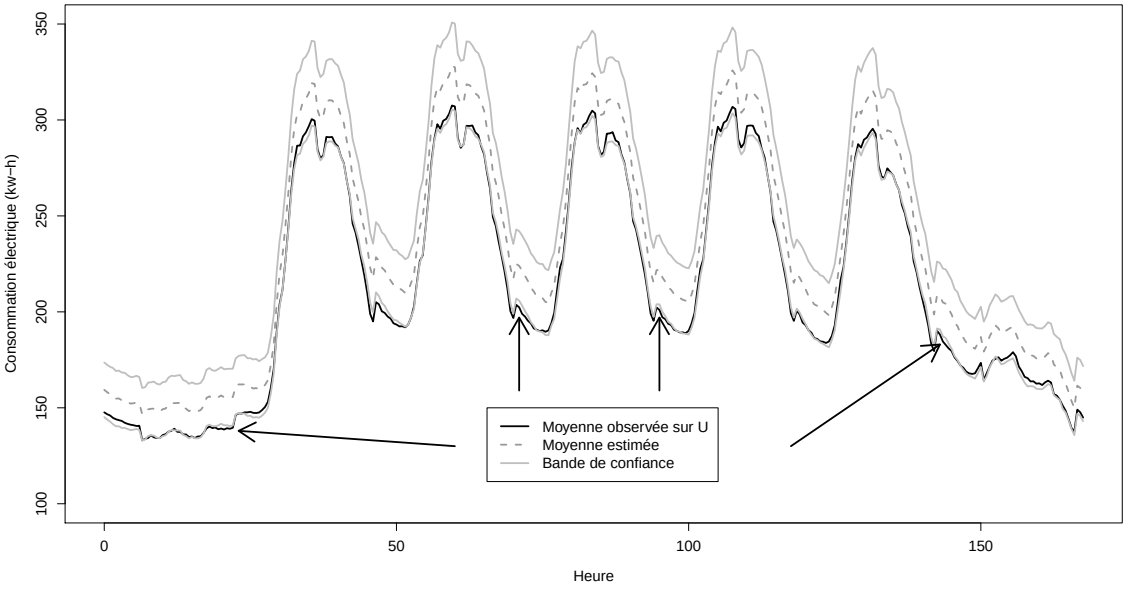
Un autre indicateur intéressant est la largeur moyenne de la bande de confiance,

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

dont les valeurs sont présentées dans la Table 6. Les deux méthodes fournissent des bandes de confiance dont les largeurs sont similaires. On note également que l'utilisation de la



(A) La courbe moyenne observée appartient à la bande de confiance



(B) La courbe moyenne observée n'appartient pas à la bande de confiance aux instants indiqués par les flèches

FIGURE 2 – Exemples de bande de confiance

variable auxiliaire permet de diminuer sensiblement la largeur moyenne des bandes, celle-ci est divisée par 2 si on considère un des plans stratifiés plutôt qu'un plan SRSWOR.

Méthodes	Nombre M de processus	Bootstrap		Processus Gaussien	
		α=0.05	α=0.01	α=0.05	α=0.01
SRSWOR	5000	35.98	43.35	35.99	43.19
STRAT (1)	5000	16.64	18.92	16.62	18.88
STRAT (2)	5000	17.58	19.99	17.55	19.94
πps	1000	17.85	20.31	17.62	19.93
MA	5000	20.03	22.93	19.72	22.40

TABLE 6 – Largeur moyenne des bandes de confiance, pour $I = 2000$ réplications.

Les Figures 3 et 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 3). A l'opposé, pour le plan stratifié (Figure 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.

5.4 Temps de calcul

Les temps de calcul avec la méthode par bootstrap sont largement supérieurs, de l'ordre d'un facteur de 1 à 1000, à ceux de la méthode par simulations de processus gaussiens (cf. Table 7). 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 SRSWOR même si utilisés individuellement leur temps de calcul reste tout à fait raisonnable.

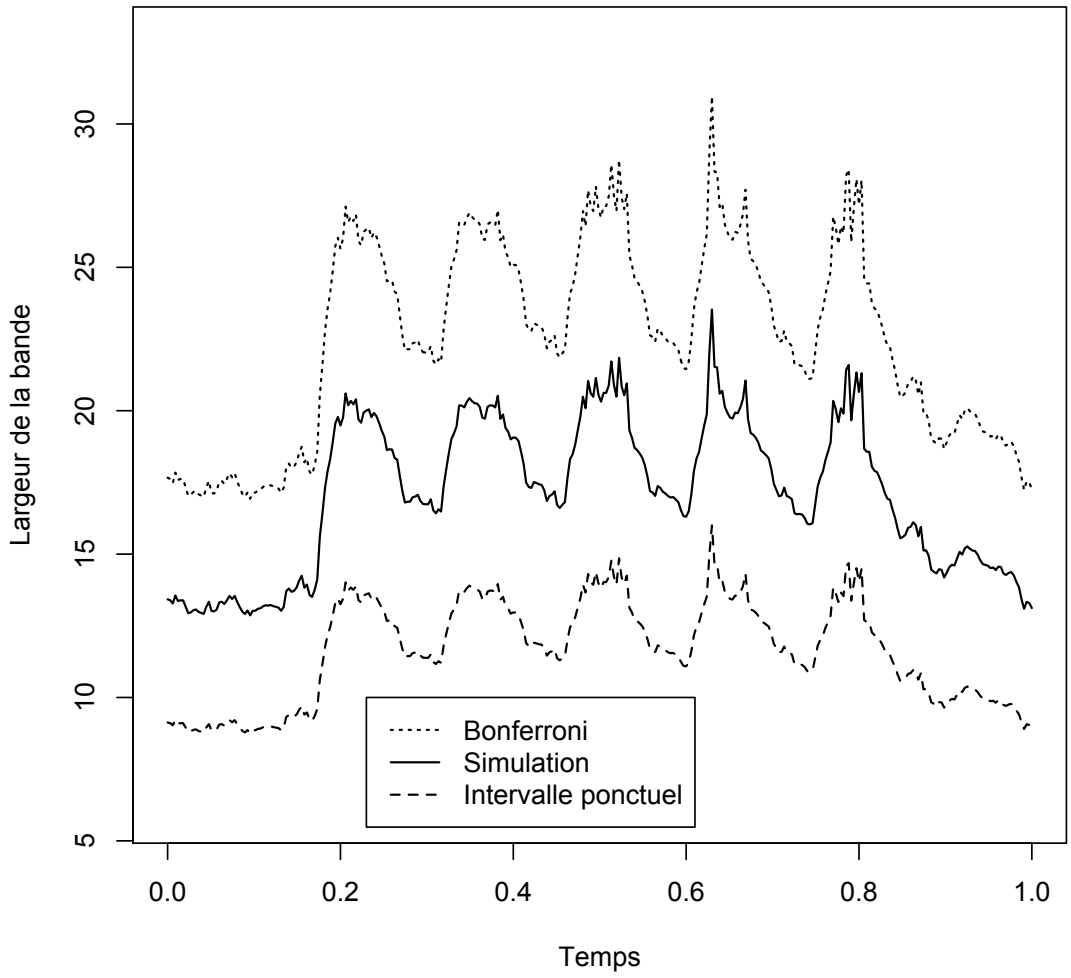


FIGURE 3 – Sondage aléatoire simple sans remise. Largeur des bandes de confiance ponctuelles, globales par simulations de processus et avec Bonferroni ($\alpha = 0.05$).

6 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

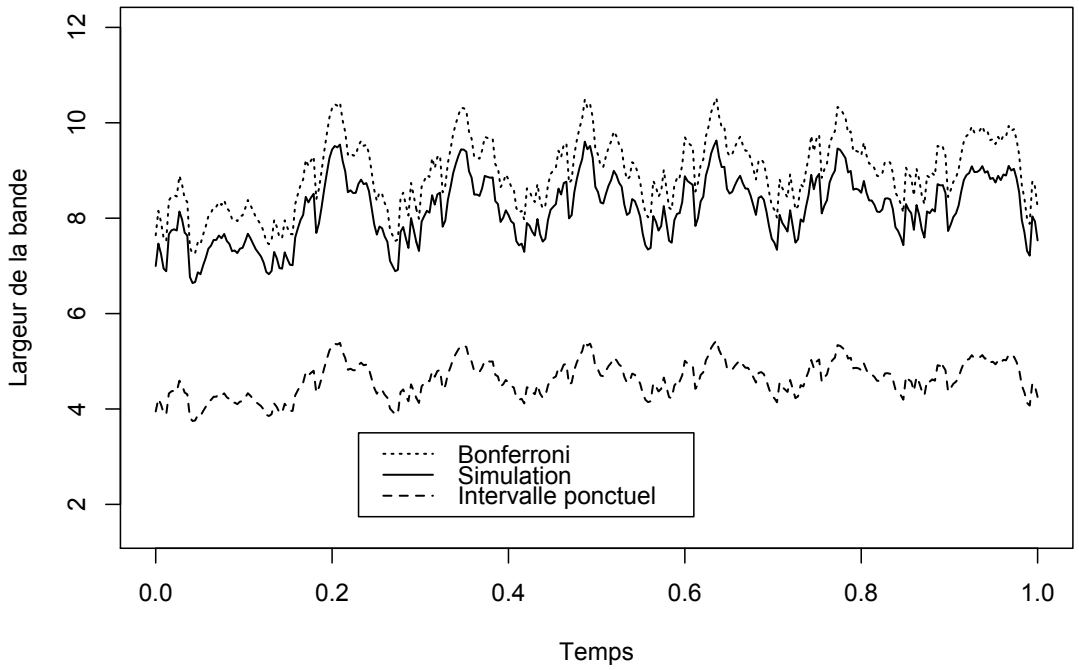


FIGURE 4 – Sondage stratifié (STRAT 1). Largeur des bandes de confiance ponctuelles, globales par simulations de processus et avec Bonferroni (avec $\alpha = 0.05$).

par rapport à un sondage de type SRSWOR.

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.

Certaines techniques suggérées dans ce travail méthodologique restent cependant sans justification asymptotique rigoureuse. Il s'agit en particulier de l'approche assistée par un modèle, pas nécessairement linéaire, à réponse fonctionnelle ainsi que des plans de sondage de type πps dont l'approximation asymptotique de la variance par la formule de Hájek doit être vérifiée. Ces questions sont l'objet de travaux en cours.

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

Stratégie	Bootstrap	Processus Gaussiens
SRSWOR	1170.6	1.0
STRAT	1839.5	1.4
πps	5020.0	7.3
MA	2423.4	1.4

TABLE 7 – Temps d'exécution d'une simulation en secondes pour M=5000 réplications. Les stratégies SRSWOR, MA et STRAT ont été programmés avec R et πps avec SAS.

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 d'estimateurs consistants de la fonction de covariance.

Remerciements : Nous remercions Guillaume Chauvet et Jean-Claude Deville pour leurs remarques fructueuses qui ont permis d'améliorer ce travail.

Références

Bickel, P. and 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., and Hall, P. (1994). Bootstrap methods for finite population. *Journal of the American Statistical Association*, 89 :75–91.

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

Cardot, H., Degras, D., and Josserand, E. (2011). Confidence bands for horvitz-thompson estimators using sampled noisy functional data. Technical report, arxiv.org 1105.2135. Revised for *Bernoulli*.

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

Chaouch, M. and Goga, C. (to appear 2012). Using complex surveys to estimate the l_1-median of a functional variable : application to electricity load curves. *International Statistical Review*.

Chauvet, G. (2007). *Méthodes de bootstrap en population finie*. PhD thesis, Université de Rennes II.

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

Cochran, W. (1977). *Sampling techniques.* John Wiley and sons, New York, 3rd edition.

Cuevas, A., Febrero, M., and 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. and Pousse, A. (1976). *Les analyse factorielles en calcul des probabilités et en statistique : essai d'étude synthétique.* PhD thesis, 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. In Guibert, P., Haziza, D., Ruiz-Gazen, A., and Tillé, Y., editors, *Méthodes de Sondages*, pages 353–357. Dunod, France.

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

Deville, J. and Tillé, Y. (2004). Efficient balanced sampling : the cube algorithm. *Biometrika*, 91 :893–912.

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

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman and Hall/CRC.

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

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

Goga, C. and Ruiz-Gazen, A. (2012). Efficient estimation of nonlinear finite population parameters using nonparametrics. preprint.

Gross, S. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*, pages 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. In Rao, C. and Pfeffermann, D., editors, *Sample Surveys : Theory Methods and Inference*, volume 29 of *Handbook of Statistics*, pages 215–246. North-Holland.

Helmers, R. and Wegkamp, M. (1998). Wild bootstrapping in finite population with auxiliary information. *Scandinavian Journal of Statistics*, 25 :383–399.

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

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Journal of the American Statistical Association*, 21 :255–285.

Ramsay, J. and Silverman, B. (2005). *Functional data analysis*. Springer, New York, second edition.

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

Särndal, C. E., Swensson, B., and 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). Ten years of balanced sampling with the cube method : an appraisal. *Survey Methodology*, 37 :215–226.

Yates, F. and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *J. Royal Statist. Soc., B*, 15 :235–261.