Sondages pour données fonctionnelles : construction de bandes de confiance asymptotiques et prise en compte d'information auxiliaire

Etienne Josserand

UNIVERSITÉ DE BOURGOGNE
U.F.R. Sciences et Techniques
Institut de Mathématiques de Bourgogne
UMR 5584 du CNRS

THÈSE

pour l'obtention du grade de

Docteur de l'Université de Bourgogne

Discipline : Mathématiques appliquées

par

Etienne Josserand

le 12 octobre 2011

Sondages pour données fonctionnelles : construction de bandes de confiance asymptotiques et prise en compte d'information auxiliaire

Directeur de thèse : **Hervé Cardot**

Membres du jury :

Frédéric Ferraty	Université Paul Sabatier	Rapporteur
Jean Opsomer	Colorado State University	Rapporteur
Jean-Claude Deville	CREST - ENSAI	Examinateur
Camelia Goga	Université de Bourgogne	Examinatrice
Anne Ruiz-Gazen	Université Toulouse Capitole	Examinatrice
Alain Dessertaine	La Poste - Pôle Expertise	Invité
Hervé Cardot	Université de Bourgogne	Directeur

Bourgogne Conseil régional

Résumé

Lorsque des bases de données fonctionnelles sont trop grandes pour être observées de manière exhaustive, les techniques d'échantillonnage fournissent une solution efficace pour estimer des quantités globales simples, telles que la courbe moyenne, sans être obligé de stocker toutes les données. Dans cette thèse, nous proposons un estimateur d'Horvitz-Thompson de la courbe moyenne, et grâce à des hypothèses asymptotiques sur le plan de sondage nous avons établi un Théorème Central Limite Fonctionnel dans le cadre des fonctions continues afin d'obtenir des bandes de confiance asymptotiques.

Pour un plan d'échantillonnage à taille fixe, nous montrons que le sondage stratifié peut grandement améliorer l'estimation comparativement au sondage aléatoire simple. De plus, nous étendons la règle d'allocation optimale de Neyman dans le contexte fonctionnel. La prise en compte d'information auxiliaire a été développée grâce à des estimateurs par modèle assisté, mais aussi en utilisant directement cette information dans les poids d'échantillonnage avec le sondage à probabilités inégales proportionnelles à la taille.

Le cas des courbes bruitées est également étudié avec la mise en place d'un lissage par polynômes locaux. Pour sélectionner la largeur de la fenêtre de lissage, nous proposons une méthode de validation croisée qui tient compte des poids de sondage. Les propriétés de consistance de nos estimateurs sont établies, ainsi que la normalité asymptotique des estimateurs de la courbe moyenne.

Deux méthodes de constructions des bandes de confiance sont proposées. La première utilise la normalité asymptotique de nos estimateurs en simulant un processus Gaussien conditionnellement à sa fonction de covariance afin d'en estimer la loi du sup. La seconde utilise des techniques de bootstrap en population finie qui ne nécessitent pas l'estimation de la fonction de covariance.

Mots clés : Données fonctionnelles, Échantillonnage, Théorème Central Limite Fonctionnel, Supremum de processus Gaussiens, Estimateur d'Horvitz-Thompson, Estimateurs par modèle assisté, Bandes de confiance asymptotiques, Bootstrap.

Abstract

When collections of functional data are too large to be exhaustively observed, survey sampling techniques provide an effective way to estimate global quantities such as the population mean function, without being obligated to store all the data. In this thesis, we propose a Horvitz–Thompson estimator of the mean trajectory, and with additional assumptions on the sampling design, we state a functional Central Limit Theorem and deduce asymptotic confidence bands.

For a fixed sample size, we show that stratified sampling can greatly improve the estimation compared to simple random sampling. In addition, we extend Neyman's rule of optimal allocation to the functional context. Taking into account auxiliary information has been developed with model-assisted estimators and weighted estimators with unequal probability sampling proportional to size.

The case of noisy curves is also studied with the help local polynomial smoothers. To select the bandwidth, we propose a cross-validation criterion that takes into account the sampling weights. The consistency properties of our estimators are established, as well as asymptotic normality of the estimators of the mean curve.

Two methods to build confidence bands are proposed. The first uses the asymptotic normality of our estimators by simulating a Gaussian process given estimated the covariance function in order to estimate the law of supremum. The second uses bootstrap techniques in a finite population that does not require to estimate the covariance function.

Keywords: Functional data, Survey sampling, Functional Central Limit Theorem, Supremum of Gaussian processes, Horvitz-Thompson estimator, Model assisted estimator, Asymptotic confidence bands, Bootstrap.

Table des matières

Chapitre 1

Introduction

L'analyse de données fonctionnelles généralise la statistique multidimensionnelle en considérant les unités statistiques comme des courbes ou des fonctions. Les motivations autour de ce sujet restent les mêmes qu'en statistique classique : décrire aux mieux les données, construire des modèles statistiques d'un ensemble de courbes, ou encore faire de la prédiction. Les premiers travaux remontent aux années 1970 (Deville, 1974, et Dauxois & Pousse, 1976) et portent essentiellement sur la généralisation de l'Analyse en Composantes Principales (ACP) dans le cas de variables Hilbertiennes. Grâce l'augmentation des performances des ordinateurs et des capacités de stockage, de nombreuses publications sont apparues dans différents domaines comme en économie (Kneip & Utikal, 2001), climatologie (Besse *et al.*, 2000), télédétection (Cardot *et al.*, 2003), bio-informatique (Mueller *et al.*, 2006), ou en marketing (Sood, James & Tellis, 2009). Les livres de Ramsay & Silverman (2002 et 2005) présentent différentes méthodes d'analyse de données fonctionnelles avec leur mise en pratique sur de nombreux exemples. On s'intéresse notamment à modéliser les données avec différents modèles de régression : linéaire (Cardot, Ferraty et Sarda, 1999), de type linéaire PLS (Partial Least Square) qui est une méthode de régression de variables réponses en fonction de variables prédictives (Preda et Saporta, 2002), linéaire généralisée (Müller & Stadtmüller, 2005), non-linéaire par noyaux (Ferraty & Vieu, 2010). Le livre de Ferraty & Vieu (2006) effectue un large panorama des méthodes nonparamétrique dans l'analyse de données fonctionnelles avec quelques illustrations pratiques. Dans la pratique, on n'a pas forcément accès à des courbes mais plutôt à des courbes évaluées en certains points au cours du temps. Les méthodes de lissage *via* des noyaux, des splines, ou encore polynômes locaux font également leur apparition pour prendre en compte la discrétisation des données. Cardot *et. al* (2003), par exemple, étendent le modèle linéaire généralisé en utilisant des splines. La description des données est également un thème de recherche actif avec la généralisation au cas fonctionnel des méthodes comme l'ACP, qui cherche à représenter au mieux la variabilité des données sur une nouvelle base de fonctions. Hall (2010) réalise un tour d'horizon des techniques d'ACP fonctionnelle en mettant en évidence les limites et perspectives de ces méthodes. Les données *sparse* fonctionnelles furent d'abord étudiées dans le cadre de l'ACP par Rice & Wu (2001), puis des méthodes plus générales furent développées par la suite par Yao *et al.* (2005) et James (2010). Hall, Müller et Wang (2006) établirent le même genre de résultats avec seulement quelques points de mesures, à condition que ces points de mesures soient aléatoires. Des méthodes de classification de courbes sont également développées dans le cadre des données fonctionnelles,

Baíllo *et al.* (2010) en font un récapitulatif avec notamment la classification supervisée et non-supervisée, illustrées par quelques exemples pratiques.

De nos jours, grâce à l'acquisition automatique de données au cours du temps, on a potentiellement accès à d'importantes base de données de données fonctionnelles. C'est notamment le cas d'E.D.F. (Électricité De France) qui prévoit d'installer dans tous les foyers et entreprises des compteurs dits *communicants*, capables d'envoyer la consommation électrique mesurée sur de petites échelles de temps (pas 10 ou 30 minutes par exemple). Il sera alors possible de connaître la consommation moyenne ou totale instantanée de parties du réseau électrique (régions de France, types de clients, ...), ce qui n'est pas le cas à l'heure actuelle. Néanmoins, pour des raisons de coût de stockage ou de bande passante, il ne sera sans doute pas possible de récupérer toutes les courbes de consommation électrique puis de les analyser. Chiky (2009) compare alors deux approches pour estimer la courbe moyenne. La première consiste en un sondage exhaustif avec technique de compression du signal en ligne (segmentation adaptative, ondelettes, *etc*). La seconde est une approche type sondage qui permet de récupérer les trajectoires complètes pour un échantillon sélectionné d'appareils. La conclusion est que des plans de sondage (même très simples) fonctionnent souvent bien mieux pour estimer un indicateur global, comme la courbe moyenne ou totale.

Cette thèse a pour objectif de formaliser cette idée, en introduisant la théorie des sondages dans le cadre d'analyse de données fonctionnelles, avec des questions simples mais nouvelles. A savoir, estimer la précision de la courbe moyenne et construire des bandes de confiance de celle-ci. Une première idée possible serait de construire des boules de confiance pour la courbe moyenne dans l'espace L^2. Mas (2007) exploite cette idée dans un test statistique basé sur la moyenne d'un échantillon fonctionnel et régularisé par l'opérateur de covariance inverse. Bunea *et al.* (2011) quant à eux construisent des régions de confiance conservatives de la moyenne d'un processus Gaussien en utilisant des estimateurs de projection adaptative. Nous adoptons dans cette thèse un point de vue différent, qui consistera de reprendre des résultats obtenus dans l'espace des fonctions continues C équipé de la norme sup. Certains de ces résultats ont été obtenus récemment par Degras (2009). Ceci permet de construire des bandes de confiance qui peuvent être visualisées et interprétées, contrairement aux boules de confiance dans un espace L^2. A partir d'un échantillon de courbes comme sur la figure 1.1, on souhaiterait construire un estimateur de la courbe moyenne avec des bandes de confiance comme illustré sur la figure 1.2.

1.1 Rappels sur la Théorie des Sondages

La théorie des sondages constitue un ensemble d'outils statistiques permettant d'étudier une population finie *via* une partie de celle-ci, appelée échantillon. Celui-ci peut être sélectionné selon un plan de sondage qui est aléatoire. On définit pour cela une loi de probabilité (discrète) sur l'ensemble des parties de la population. Un des sondages les plus simples et qui sera détaillé par la suite est le sondage aléatoire simple sans remise. Pour une taille d'échantillon n fixée, tous les ensembles constitués de n éléments distincts ont la même probabilité d'être sélectionnés. Comme le plan est de taille fixe, la condition *i.i.d.* sur les individus, souvent supposée en statistique inférentielle classique, n'est alors plus satisfaite. On peut généraliser cette remarque à la majorité des plans de sondage et c'est tout l'intérêt de la théorie des sondages de développer des techniques d'estimation dans

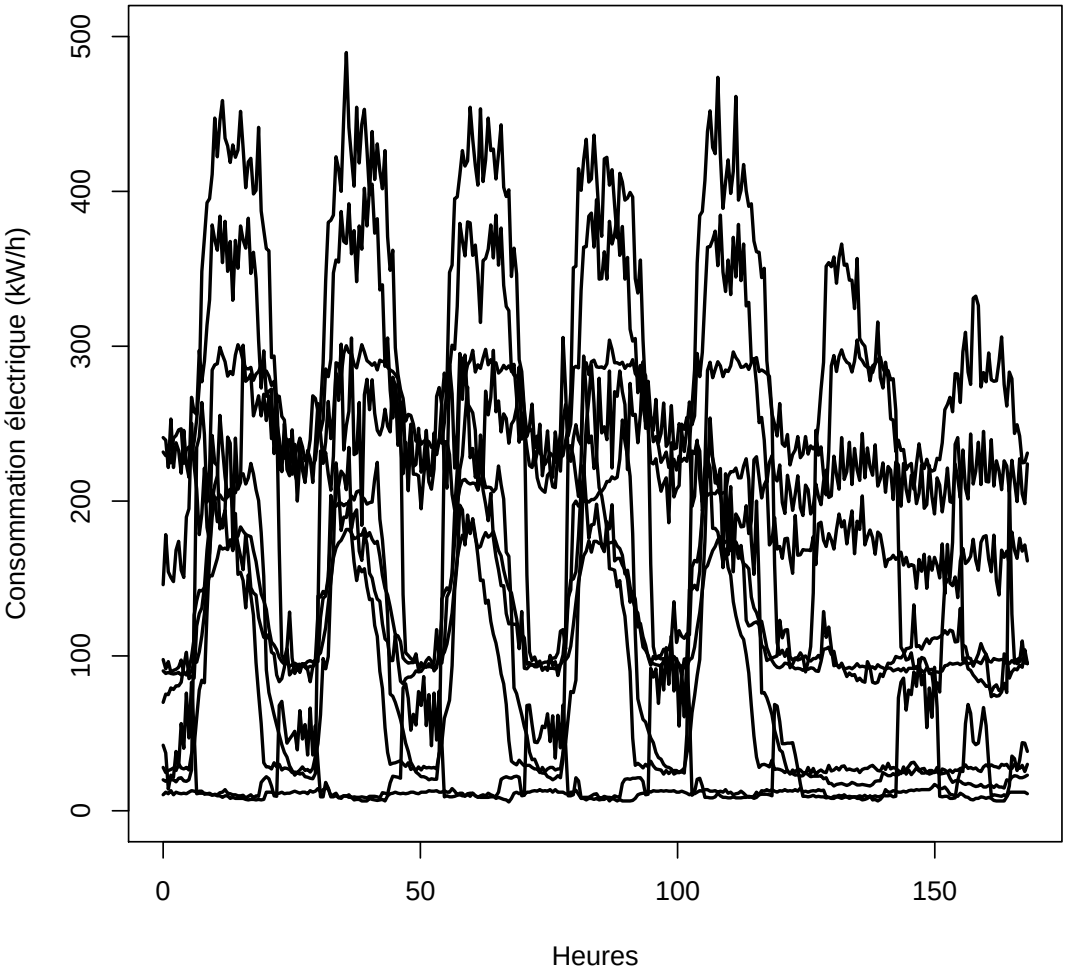

FIGURE 1.1 – Echantillon de 10 courbes de consommation électrique.

un tel cadre.

Nous allons présenter quelques généralités sur la théorie des sondages qui sont à la base des travaux présentés aux chapitres suivants. Pour de plus amples détails, on pourra se référer aux ouvrages de référence que sont les livres de Särndal *et al.* (1992) et Tillé (2001).

1.1.1 Notations

On considère une population $U = \{1, \ldots, k, \ldots, N\}$ de taille finie N et on s'intéresse à une variable d'intérêt Y définie pour chaque individu k de la population U. Pour chaque unité k, on note Y_k la valeur déterministe du caractère Y associée. En théorie des sondages,

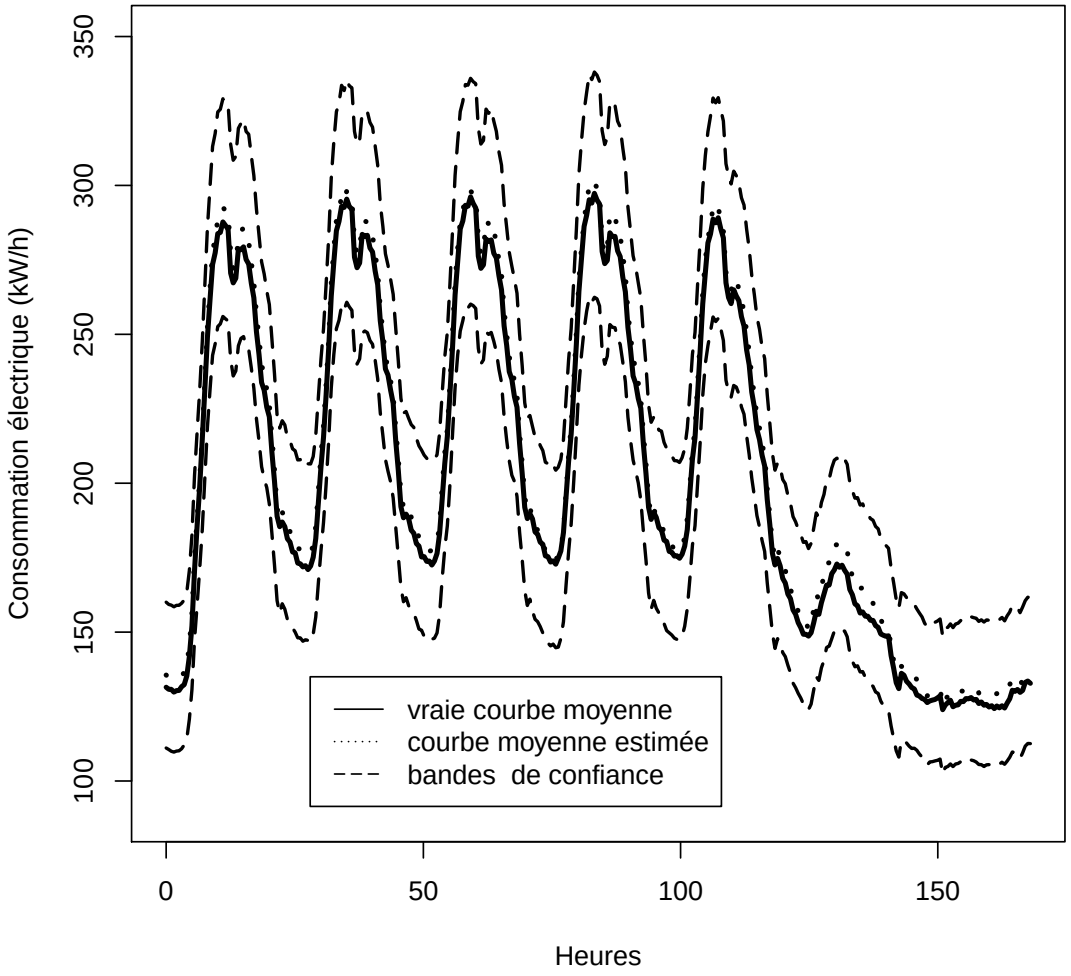

FIGURE 1.2 – Exemple de construction de bandes de confiance.

on s'intéresse à une fonction de la variable d'intérêt

$$\theta \quad = \quad \theta(Y_k, k \in U) \tag{1.1}$$

appelée paramètre d'intérêt ou fonctionnelle, que l'on cherche à estimer le plus précisément possible en utilisant un échantillon. Les paramètres d'intérêts de la variable Y les plus souvent étudiés sont notamment :

– le total

$$t_Y \quad = \quad \sum_{k \in U} Y_k, \tag{1.2}$$

– la moyenne

$$\mu_Y = \frac{1}{N}\sum_{k\in U} Y_k, \tag{1.3}$$

– la variance

$$\sigma_Y^2 = \frac{1}{N}\sum_{k\in U}(Y_k - \mu_Y)^2. \tag{1.4}$$

Récupérer toutes les valeurs de Y sur toute la population U, c'est-à-dire réaliser un sondage exhaustif, n'est pas toujours possible. En effet, les contraintes de temps ou de coût peuvent être importantes, d'autant plus si la taille de la population N est grande. On peut alors considérer seulement une partie de la population.

Un échantillon s, *i.e.* une partie $s \subset U$, est tiré selon un procédé probabiliste $p(s)$ où p est une loi de probabilité sur l'ensemble de parties possibles de U. Le plan de sondage p vérifie

$$\forall s \subset U \quad p(s) \geq 0 \text{ et } \sum_{s\subset U} p(s) = 1. \tag{1.5}$$

L'échantillon aléatoire est noté S et sa taille n_S qui peut être aussi aléatoire. Par la suite, nous étudierons principalement des plans de sondage à taille fixe et noterons simplement n la taille de l'échantillon.

1.1.2 Estimateurs d'Horvitz-Thompson

En reprenant les notations de Särndal *et al* (1992), on note $I_k = \mathbb{1}_{k\in s}$ l'indicatrice d'appartenance à l'échantillon de l'unité k. On appelle probabilité d'inclusion d'ordre 1 de l'individu k la probabilité π_k de cet individu d'être sélectionné dans l'échantillon s, $\pi_k = \mathbb{P}(k \in s)$. De façon analogue, on appelle probabilité d'inclusion d'ordre 2 des individus k, l la probabilité π_{kl} qu'ont ces individus d'être sélectionnés dans l'échantillon s, $\pi_{kl} = \mathbb{P}(k, l \in s)$. Par la suite, on supposera toujours que $\pi_k > 0$ et $\pi_{kl} > 0$ pour toutes les unités k et l de U. Ce qui signifie que chaque individu k et chaque couple d'individus k, l ont une probabilité non nulle d'être sélectionnés.

Théorème 1.1.1 (Horvitz and Thompson (1952))**.** *Si pour toute unité k de U on a $\pi_k > 0$, alors*

$$\widehat{t}_{Y\pi} = \sum_{k\in s}\frac{Y_k}{\pi_k} = \sum_{k\in U} Y_k \frac{I_k}{\pi_k}$$

est un estimateur sans biais de t_Y.

L'estimateur $\widehat{t}_{Y\pi}$ de t_Y est appelé estimateur d'Horvitz-Thompson ou encore π-estimateur du total t_Y. Son principal intérêt est que son expression reste simple, c'est une moyenne pondérée par des poids $1/\pi_k$, et c'est un estimateur sans biais puisque $\mathbb{E}(I_k) = \pi_k$. C'est par ailleurs le seul estimateur linéaire sans biais d'un total (Tillé, 2001, section 3.14). On peut remarquer que la valeur de $\widehat{t}_{Y\pi}$ dépend directement des individus k sélectionnés dans l'échantillon s, et que la partie aléatoire de l'estimateur repose sur la sélection ou la non-sélection d'un individu.

Afin de mesurer la précision de cet estimateur, on peut calculer sa variance.

Théorème 1.1.2 (Horvitz and Thompson (1952))**.** *La variance de l'estimateur du total* $\widehat{t}_{Y\pi}$ *est*

$$V(\widehat{t}_{Y\pi}) = \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}$$

où $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ *et* $\Delta_{kk} = \pi_k(1-\pi_k)$*. L'estimateur de variance de Horvitz-Thompson est donné par*

$$\widehat{V}(\widehat{t}_{Y\pi}) = \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}.$$

Lorsque le plan est de taille n fixe, Sen (1953) et Yates and Grundy (1953) ont montré que la variance pouvait être formulée d'une façon particulière. Ce qui conduit à un deuxième estimateur de variance.

Théorème 1.1.3 (Sen (1953) ; Yates and Grundy (1953))**.** *Si le plan est de taille fixe, la variance de l'estimateur du total* $\widehat{t}_{Y\pi}$ *est*

$$V(\widehat{t}_{Y\pi}) = -\frac{1}{2}\sum_{k\in U}\sum_{l\neq k\in U}\Delta_{kl}\left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2$$

où $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ *et* $\Delta_{kk} = \pi_k(1-\pi_k)$*. Cette variance peut être estimée sans biais par l'estimateur de Sen-Yates-Grundy*

$$\widehat{V}(\widehat{t}_{Y\pi}) = -\frac{1}{2}\sum_{k\in s}\sum_{l\neq k\in s}\frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2.$$

L'inconvénient de cet estimateur est qu'il peut prendre des valeurs négatives. Une condition pour qu'il soit toujours positif est que $\pi_k\pi_l - \pi_{kl} \geq 0$ pour tout $k, l \in U$ avec $k \neq l$. Cette condition est appelée condition de Sen-Yates-Grundy. Elle est notamment vérifiée par le sondage aléatoire simple sans remise et le sondage stratifié que nous allons maintenant présenter.

1.1.3 Plans de sondage utilisés

Nous présentons ici trois plans de sondage classiques mis en œuvre au cours de la thèse :

- plan simple sans remise,
- plan stratifié,
- plan à probabilités inégales proportionnelles à la taille,

qui seront utilisés dans les chapitres suivants. La taille N de la population U est supposée connue par la suite.

Plan simple sans remise

Lorsque N est connu et la taille de l'échantillon n est fixée, on peut effectuer un sondage aléatoire simple sans remise. Sur l'ensemble de tous les échantillons de taille fixe n, on dit qu'un plan est simple si tous les échantillons ont la même probabilité d'être sélectionnés.

Étant donné qu'il y a C_N^n échantillons différents de taille n , on peut expliciter le plan de sondage aléatoire simple ainsi

$$p(s) = \begin{cases} 1/C_N^n & \text{si } \#(s) = n \\ 0 & \text{sinon} \end{cases} . \tag{1.6}$$

Les probabilités d'inclusion du premier et du second ordre peuvent être facilement calculées,

$$\pi_k = \frac{C_{N-1}^{n-1}}{C_N^n} = \frac{n}{N}, \quad k \in U, \tag{1.7}$$

$$\pi_{kl} = \frac{C_{N-2}^{n-2}}{C_N^n} = \frac{n(n-1)}{N(N-1)}, \quad k, l \in U.\ k \neq l. \tag{1.8}$$

L'estimateur d'Horvitz-Thompson $\widehat{\mu}_{Y\pi}$ de la moyenne μ_Y devient simplement

$$\widehat{\mu}_{Y\pi} = \frac{1}{n} \sum_{k \in s} Y_k \tag{1.9}$$

et sa variance est donnée par

$$V(\widehat{\mu}_{Y\pi}) = (1-f)\frac{S_Y^2}{n} \tag{1.10}$$

où $f = n/N$ représente le taux de sondage et S_Y^2 représente la variance corrigée de Y. Plus le taux de sondage augmente, plus on se rapproche de l'exhaustivité et donc la variance $V(\widehat{\mu}_{Y\pi})$ diminue.

Plan stratifié et allocation optimale

Supposons que la population U soit partitionnée en H sous-ensembles, U_h, $h = 1, \ldots, H$, appelés strates telles que

$$\bigcup_{h=1}^{H} U_h = U \text{ et } U_i \bigcap U_j = \emptyset \text{ pour } i \neq j. \tag{1.11}$$

Les strates sont généralement construites à partir d'une information auxiliaire, souvent qualitative, connue sur la population entière (catégorie socioprofessionnelle, type de contrat,...). Le nombre d'éléments N_h de la strate U_h est appelé taille de la strate. On a $\sum_{h=1}^{H} N_h = N$. Les N_h sont supposés connus et constituent une information auxiliaire disponible sur la population entière. La moyenne sur U_h est donnée par

$$\mu_h = \frac{1}{N_h} \sum_{k \in U_h} Y_k. \tag{1.12}$$

De plus, on note σ_h^2 la variance de Y dans la strate h et S_h^2 sa variance corrigée

$$\sigma_h^2 = \frac{1}{N_h} \sum_{k \in U_h} (Y_k - \mu_h)^2 , \tag{1.13}$$

$$S_h^2 = \frac{N_h}{N_h - 1}\sigma_h^2. \tag{1.14}$$

En sondage stratifié avec sondage aléatoire simple sans remise dans toutes les strates, les probabilités d'inclusion de premier et de second ordre sont explicitement connues,

$$\begin{aligned}
\pi_k &= \frac{n_h}{N_h}, \quad k \in U_h, \\
\pi_{kl} &= \frac{n_h(n_h-1)}{N_h(N_h-1)}, \quad k \text{ et } l \in U_h, \\
\pi_{kl} &= \frac{n_h n_i}{N_h N_i}, \quad k \in U_h \text{ et } l \in U_i,
\end{aligned}$$

où $n_h,\ n_h < N_h$, est le nombre d'unités sélectionnées dans la strate h. L'estimateur stratifié de la moyenne $\widehat{\mu}_{\text{strat}}$ de μ_Y est défini par

$$\widehat{\mu}_{\text{strat}} = \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k\in s_h} Y_k, \tag{1.15}$$

où s_h est un échantillon de taille n_h obtenu par un sondage aléatoire simple sans remise dans la strate U_h. Il est facile de montrer que

$$V(\widehat{\mu}_{\text{strat}}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h}\ S_h^2. \tag{1.16}$$

La variance dépend alors directement du choix de la taille de l'échantillon n_h dans chaque strate h. Par exemple, on dit qu'un plan stratifié est à allocation proportionnelle si

$$\frac{n_h}{N_h} = \frac{n}{N},\ h = 1, \ldots, H. \tag{1.17}$$

Si on suppose qu'il est possible de calculer des $n_h = nN_h/N$ entiers, alors on peut définir l'estimateur à allocation proportionnelle de la moyenne

$$\widehat{\mu}_{\text{prop}} = \frac{1}{n}\sum_{k\in s} Y_k, \tag{1.18}$$

et sa variance est

$$V(\widehat{\mu}_{\text{prop}}) = \frac{N-n}{nN^2}\sum_{h=1} H N_h S_h^2. \tag{1.19}$$

Néanmoins, l'allocation proportionnelle n'est en général pas le meilleur choix dans le cas d'un sondage stratifié si l'on cherche à estimer un total ou une moyenne. En effet, on souhaite déterminer une allocation optimale des strates, qui permet d'obtenir un estimateur de la moyenne de variance minimale, pour une taille globale de l'échantillon fixée.

Proposition 1.1.1 (Neyman (1934))**.** *La solution du problème d'allocation optimale*

$$\min_{(n_1,\ldots,n_H)} V\big(\widehat{\mu}_{strat}\big) \quad \textit{avec } \sum_{h=1}^{H} n_h = n \textit{ et } n_h > 0,\ h = 1, \ldots, H.$$

est donnée par

$$n_h^* = n\ \frac{N_h S_h}{\sum_{i=1}^{H} N_i S_h}, \quad h = 1, \ldots, H. \tag{1.20}$$

On notera $\widehat{\mu}_{\text{optim}}$ l'estimateur de la moyenne ayant pour allocation les n_h^* pour $h = 1, \ldots, H$. Il est possible de mesurer le gain obtenu entre une allocation optimale et une allocation proportionnelle. En effet, on peut montrer que

$$V(\widehat{\mu}_{\text{optim}}) - V(\widehat{\mu}_{\text{prop}}) = \frac{1}{n} \sum_{h=1}^{H} \frac{N_h}{N} \left[S_h - \left(\sum_{h=1}^{H} \frac{N_h}{N} S_h \right) \right]^2 . \tag{1.21}$$

Le gain en précision entre les deux estimateurs dépend directement de la variance des écarts-types des strates. Ainsi, plus les écarts-types des strates seront hétérogènes, meilleur sera l'estimateur avec allocation optimale.

Dans le contexte fonctionnel que l'on développera par la suite, et comme dans le cas multivarié (voir *e.g.* Cochran, 1977), déterminer une allocation optimale dépend clairement du critère à minimiser (Chapitre 3).

Plan à probabilités inégales proportionnelles à la taille

Lorsqu'aucune information n'est disponible sur la population, on utilise généralement des plans simples. Mais si l'on dispose d'une variable auxiliaire connue préalablement sur toute la population, il peut être intéressant d'introduire un plan à probabilités inégales qui tient compte de cette information.

On suppose disposer d'un caractère auxiliaire positif, $x_k > 0$, connu pour toutes les unités k de la population U. De plus, si on suppose que le caractère x est *quasi* proportionnel à la variable d'intérêt Y, il est intéressant de sélectionner les unités avec des probabilités d'inclusions π_k proportionnelles au caractère auxiliaire x_k. On remarque en effet que si le plan est de taille fixe, la variance du π-estimateur de la moyenne est donnée dans le théorème 1.1.3 par

$$V(\widehat{\mu}_{Y\pi}) = -\frac{1}{N^2} \frac{1}{2} \sum_{k \in U} \sum_{l \neq k \in U} \Delta_{kl} \left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l} \right)^2 . \tag{1.22}$$

Si les probabilités d'inclusions π_k sont approximativement proportionnelles aux Y_k, alors les Y_k/π_k, $k \in U$, deviennent approximativement constants et donc la variance $V(\widehat{\mu}_{Y\pi})$ est proche de zéro. L'intérêt du plan à probabilités inégales sera donc très important lorsqu'il y a un lien proportionnel entre la variable d'intérêt Y et le caractère auxiliaire x.

Les probabilités d'inclusions du premier ordre, pour un plan à taille fixe n, sont données par

$$\pi_k = n \frac{x_k}{\sum_{k \in U} x_k}, \text{ pour tout } k \in U. \tag{1.23}$$

On note que les π_k obtenus peuvent être supérieurs à un. Pour remédier à ce problème, les $\pi_k > 1$ sont fixés à un et les unités correspondantes sont automatiquement sélectionnées. Pour les unités restantes, on recalcule les probabilités avec la formule (1.23) sans prendre en compte les unités déjà sélectionnées. On itère cette procédure jusqu'à ce que tous les π_k, $k \in U$, soient inférieurs ou égaux à un (Tillé, 2001).

Le π-estimateur de la moyenne respectant les probabilités d'inclusions définies en (1.23) est noté $\widehat{\mu}_{\pi ps}$. Celui-ci n'est pas unique puisqu'il existe une infinité de plan de sondage vérifiant ces probabilités d'inclusion du premier ordre. De plus, il est difficile d'expliciter

les probabilités d'inclusions du second ordre qui sont nécessaires pour estimer la variance de l'estimateur $\widehat{\mu}_{\pi ps}$. Néanmoins, lorsque la taille du plan de sondage est fixée et le plan de sondage proche de l'entropie maximale, on peut utiliser la formule d'Hájek (1964), qui fournit des approximations avec le *rejective sampling*, pour estimer la variance (Berger, 1998),

$$\widehat{V}(\widehat{\mu}_{\pi\text{ps}}) = \frac{1}{N^2}\sum_{k\in s}(1-\pi_k)\left(\frac{Y_k}{\pi_k} - \widehat{R}\right)^2 \tag{1.24}$$

où $\widehat{R} = \sum_{k\in s}\frac{Y_k}{\pi_k}(1-\pi_k)/\sum_{k\in s}(1-\pi_k)$. Cet estimateur a la propriété d'être asymptotiquement sans biais. L'entropie d'un plan de sondage p est donnée par

$$I(p) = -\sum_{s\subset U} p(s)\log\left(p(s)\right) \tag{1.25}$$

avec la convention $0\log(0) = 0$. Plus l'entropie est élevée, plus le plan de sondage est dans un certain sens aléatoire. Ainsi l'approximation (1.24) dépend directement de l'algorithme de tirage utilisé. On pourra notamment utiliser un tirage de Poisson ou encore l'algorithme du cube (Deville & Tillé, 2000). Pour un panorama des différents algorithmes de tirage, on pourra se référer au livre de Tillé (2006).

1.1.4 Asymptotique en théorie des sondages

Comme en statistique inférentielle classique, il est intéressant de pouvoir évaluer la précision d'un estimateur. On estime souvent les variances des estimateurs afin d'en mesurer la qualité par le biais d'un intervalle de confiance. Ceci nécessite de connaître la loi asymptotique de l'estimateur. Pour cela, on doit d'abord introduire ce qu'est l'asymptotique dans le cadre de la théorie des sondages, qui n'est pas aussi naturelle qu'en statistique inférentielle classique puisque la population est de taille finie.

Modèle de superpopulation

Le modèle de superpopulation de Isaki and Fuller (1982) introduit une population limite $U_{\mathbb{N}}$ avec un nombre infini dénombrable d'unités. La variable d'intérêt Y est tirée selon une loi de probabilité et la construction de la population U_ν consiste en ν tirages indépendants. On tire ensuite une nouvelle réalisation pour obtenir la population $U_{\nu+1}$. On peut alors considérer une suite croissante de sous populations imbriquées telle que $U_1 \subset \ldots U_{\nu-1} \subset U_\nu \subset U_{\nu+1} \subset \cdots \subset U_{\mathbb{N}}$, de tailles $N_1 < N_2 < \cdots < N_\nu < \ldots$, et une suite d'échantillons s_ν de taille n_ν tirée dans U_ν selon le plan de sondage $p_\nu(s_\nu) = \mathbb{P}(s_\nu|U_\nu)$. On désigne par $\pi_{k\nu}$ et $\pi_{kl\nu}$ leur première et seconde probabilités d'inclusion. On peut d'ores et déjà remarquer que les sous-populations sont croissantes ce qui n'est pas le cas de la suite des échantillons. Pour simplifier les notations, on omet l'indice ν lorsqu'il n'y a pas d'ambiguïté.

Convergence asymptotique

Les hypothèses généralement faites dans ce cadre sont les suivantes :

(**A1**) $\lim_{N\to\infty}\frac{n}{N} = \pi \in]0,1[$.

(**A2**) $\min \pi_k \geq \lambda > 0$, $\min_{k \neq l} \pi_{kl} \geq \lambda * > 0$, $\limsup_{N \to \infty} \; n \max_{k \neq l} |\pi_{kl} - \pi_k \pi_l| < C_1 < \infty$.

L'hypothèse (A1) suppose que la taille de l'échantillon et la taille de la population augmentent à la même vitesse. L'hypothèse (A2), quant à elle, quantifie l'écart à l'indépendance des probabilités d'inclusion d'ordre 2, et permet aussi de donner la vitesse de convergence.

Pour étudier la convergence asymptotique des estimateurs, on utilise la notion de consistance.

Définition 1.1.1. *Un estimateur $\widehat{\theta}_N$ de θ_N est dit consistant si pour tout $\epsilon > 0$*

$$\lim_{N \to \infty} \mathbb{P}(|\widehat{\theta} - \theta_N| > \epsilon \mid U_N) = 0. \tag{1.26}$$

De manière similaire à Robinson & Särndal (1983), sous les hypothèses (A1) et (A2), l'estimateur d'Horvitz-Thompson de la moyenne est consistant. Robinson & Särndal utilisent en fait une hypothèse (A2) moins forte puisqu'ils supposent seulement que

$$\lim_{N \to \infty} \max_{k \neq l} \left| \frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right| = 0. \tag{1.27}$$

Breidt & Opsomer (2000) ajoute l'hypothèse

(**A3**) $\lim_{N \to \infty} \max_{(i_1,i_2,i_3,i_4) \in D_{4,N}} |E\{(I_{i_1} I_{i_2} - \pi_{i_1 i_2})(I_{i_3} I_{i_4} - \pi_{i_3 i_4})\}| = 0$,
où $D_{t,N}$ représente l'ensemble de tous les t-tuples distincts $(i_1, \ldots, i_t)$ de U_N,

pour obtenir la consistance de l'estimateur d'Horvitz-Thompson de la variance.

Toutes ces hypothèses sont pleinement satisfaites par les plans usuels comme le plan simple et le plan stratifié (Robinson & Särndal, 1983, Breidt & Opsomer, 2000).

Théorème central limite

Le théorème central limite dans le cadre sondage n'a été démontré que pour certains plans de sondage. Erdös & Rényi (1959) et Hájek (1960) ont établi le résultat pour le sondage aléatoire simple sans remise. Pour l'échantillon s_N, on note $f_N = n/N$ le taux de sondage, $\widehat{\mu}_N$ la moyenne simple du caractère Y. Pour la population U_N, on note μ_N la moyenne simple du caractère Yet S_N l'écart-type de la variable Y.

Théorème 1.1.4 (Hájek, 1960)**.** *On suppose que $n \to \infty$, et $N - n \to \infty$ quand $N \to \infty$. Alors, dans le cas du sondage aléatoire simple sans remise*

$$\sqrt{n} \; \frac{\widehat{\mu}_N - \mu_N}{\sqrt{1 - f_N} S_N} \to \mathcal{N}(0,1) \text{ quand } N \to \infty \tag{1.28}$$

si et seulement si la suite (Y_k) vérifie la condition de Lindeberg-Hájek

$$\lim_{N \to \infty} \sum_{T_N(\delta)} \frac{Y_k - \mu_N}{(N-1) S_N^2} = 0 \text{ quelque soit } \delta > 0 \tag{1.29}$$

où $T_N(\delta)$ désigne l'ensemble des unités de U pour lesquelles

$$\frac{|Y_k - \mu_\nu|}{\sqrt{1 - f_N} S_N} > \delta \sqrt{n}. \tag{1.30}$$

La normalité de l'estimateur de la moyenne avec un plan stratifié a également été étudié par Bickel & Freedman (1984) et Krewski & Rao (1981). Celle du πps par Hájek (1964) pour le plan *rejective sampling*.

Intervalles de confiance asymptotiques

En se donnant un risque d'erreur α, on peut à présent construire des intervalles de confiance. Pour le π-estimateur du total t_Y, l'intervalle de confiance est

$$I_\alpha(t_Y) = \left[\widehat{t}_{\pi Y} - z_{1-\alpha/2}\sqrt{\widehat{V}(\widehat{t}_{\pi Y})}\ ,\ \widehat{t}_{\pi Y} + z_{1-\alpha/2}\sqrt{\widehat{V}(\widehat{t}_{\pi Y})}\right] \tag{1.31}$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ d'une loi normale centrée réduite.

Bootstrap

Le bootstrap est une méthode alternative d'estimation de précision par réplication, qui permet d'éviter l'estimation de la variance de l'estimateur. Cette technique fut d'abord développée par Efron (1979) dans le cadre d'une population infinie. Gross (1980) présente une adaptation du bootstrap en population finie. On construit une population fictive U^* en répliquant les données observées. Puis, par tirages répétés dans la population U^*, on peut estimer la variance ou les quantiles d'un estimateur. Pour plus de détails on peut se référer à la thèse de Chauvet (2007) qui étudie les méthodes de bootstrap et leurs mises en œuvre sur plusieurs plans de sondage. On remarque alors que la façon de répliquer la population est directement liée au plan de sondage utilisé. Nous traiterons ici le cas de l'estimation d'un intervalle de confiance de la moyenne μ_Y.

Dans le cas d'un sondage aléatoire simple, on réplique N/n fois les n unités de l'échantillon s afin de construire la population fictive U^*. On peut ensuite tirer selon un sondage aléatoire simple, dans la population U^*, M échantillons $s_j^*, j = 1, \ldots, M$, et calculer leurs moyenne $\widehat{\mu}_{Yj}^*, j = 1, \ldots, M$. On peut alors déterminer un intervalle de confiance, grâce aux quantiles empiriques de ces M estimateurs. En effet, on note $E_{c_\alpha} = \{j|\widehat{\mu}_{Yj}^* < c_\alpha\}$ et c_α est choisi tel que

$$c_\alpha = M\alpha\#(E_{c_\alpha}). \tag{1.32}$$

On obtient alors un intervalle de confiance pour μ_Y, avec un risque d'erreur α, de la forme

$$I_\alpha(\mu_Y) = \left[c_{\alpha/2}\ ,\ c_{1-\alpha/2}\right]. \tag{1.33}$$

On peut d'ores et déjà faire deux remarques sur cet algorithme de bootstrap en population finie. La première concerne le choix de M qui doit être assez grand pour que l'estimation soit assez précise. La seconde est que N/n n'est pas nécessairement un entier. Pour palier ce problème, chaque unité est dupliquée $[N/n]$ fois, où $[.]$ désigne la fonction partie entière. On complète alors la population fictive U^* avec un tirage aléatoire simple dans s de taille $N - n[N/n]$. La population U^* sera alors potentiellement différente à chaque étape de l'algorithme.

Dans le cas d'un sondage à probabilités inégales, le principe reste le même. Chaque individu k est répliqué $1/\pi_k$ fois pour constituer la population U^*. Néanmoins, les $1/\pi_k$ n'étant pas nécessairement entiers, il faut recourir à des résolutions de gestions de l'arrondi. Chauvet (section 3.1.2, 2007) présente une méthode de bootstrap générale pour un plan à probabilités inégales :

1. Chaque unité k de s est dupliquée $[1/\pi_k]$ fois, où $[.]$ désigne la fonction partie entière. On complète les unités ainsi obtenues par un échantillon sélectionné dans s, selon le plan de sondage p avec des probabilités d'inclusions $\alpha_k = 1/\pi_k - [1/\pi_k]$. On obtient alors une population fictive U^*.

2. On échantillonne dans U^* selon le plan de sondage p (avec les probabilités d'inclusion π_k d'origine) pour obtenir l'échantillon s^*. L'estimateur $\widehat{\mu}_Y^*$ est alors calculé.
3. Les étapes 1 et 2 sont alors répétées un grand nombre M de fois dans le but d'estimer une variance ou des quantiles.

Dans le cas d'un sondage à probabilités inégales proportionnelles à une variable taille, il est nécessaire de modifier l'algorithme. En effet, le plan p est de taille fixe tandis que lors de la construction de la population fictive la quantité $\sum_{k\in U^*} \pi_k$ n'est pas nécessairement entière. Ceci peut notamment conduire à une estimation inconsistante de la variance. Dans ce cas, la solution consiste à bootstraper le processus, c'est-à-dire pour chaque unité k de U^* la probabilité

$$\pi_k^* = n\frac{x_k}{\sum_{k\in U^*}} \tag{1.34}$$

avec un recalcul si les probabilités dépassent 1. L'échantillon s^* est alors tiré selon p dans U^* en respectant les probabilités π_k^*.

1.2 Sondage sur Données Fonctionnelles

1.2.1 Estimateurs d'Horvitz-Thompson pour données fonctionnelles

Cette thèse propose d'utiliser des techniques de sondage dans le cadre de l'analyse de données fonctionnelles pour estimer une courbe moyenne ainsi que la construction de bandes de confiance. Les premiers travaux ont porté sur l'estimateur d'Horvitz–Thompson en étudiant le cas du plan de sondage stratifié. Le découpage en strates de la population permet d'améliorer la précision de l'estimateur d'Horvitz–Thompson, et la notion d'allocation optimale de type Neyman a été généralisée au cadre fonctionnel. En considérant le modèle de superpopulation de Isaki & Fuller (1982), on peut montrer la normalité asymptotique de l'estimateur de la courbe moyenne afin de construire des bandes de confiance asymptotiques. Ceci est détaillé au chapitre 2 avec une application sur une population de $N = 18902$ courbes de consommation électrique. Ce travail est publié dans le journal *Biometrika* sous la référence

- 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.

Des compléments sur les démonstrations et la stratification sont apportés au chapitre 3.

Le sondage sur données fonctionnelles consiste en l'utilisation des techniques de sondages avec une variable d'intérêt $\mathscr{Y}$ fonctionnel. En reprenant les notations précédentes, $Y_k = (Y_k(t))_{t\in[0,T]}$ devient une fonction définie sur l'intervalle $[0,T]$ que l'on suppose continue. On note $\mu \in C[0,T]$, la moyenne des Y_k de la population

$$\mu(t) = \frac{1}{N}\sum_{k\in U} Y_k(t), \quad t \in [0,T] \tag{1.35}$$

et γ, la fonction de covariance définie sur $C([0,T]\times[0,T])$ par

$$\gamma(s,t) = \frac{1}{N}\sum_{k\in U} (Y_k(s)-\mu(s))\cdot(Y_k(t)-\mu(t)), \quad (s,t)\in[0,T]\times[0,T]. \tag{1.36}$$

On suppose par la suite que la taille de la population N est connue. Les estimateurs classiques de type Horvitz-Thompson de μ et γ sont définis par

$$\widehat{\mu} = \frac{1}{N}\sum_{k\in s}\frac{Y_k}{\pi_k}, \tag{1.37}$$

$$\widehat{\gamma}(s,t) = \frac{1}{N}\sum_{k\in s}\frac{Y_k(s)\cdot Y_k(t)}{\pi_k} - \widehat{\mu}(s)\cdot\widehat{\mu}(t),\ (s,t)\in[0,T]\times[0,T]. \tag{1.38}$$

Ces estimateurs ont la propriété d'être sans biais (Cardot *et al.* 2010).

Dans la pratique, on n'observe que très rarement une courbe complète $Y_k(t)$ avec $t\in[0,T]$ mais plutôt une courbe discrétisée $Y_k(t_j)$ avec d points de discrétisations $0=t_1<\cdots<t_d=T$. Si l'on suppose qu'il n'y a pas d'erreur de mesure, ce qui dépend du problème considéré, et que les courbes sont suffisamment régulières, alors une interpolation linéaire est une solution simple et robuste pour obtenir de bonnes approximations des courbes en chaque instant t. Pour chaque unité k dans l'échantillon s, la courbe interpolée est définie par

$$\tilde{Y}_k(t) = Y_k(t_i) + \frac{Y_k(t_{i+1})-Y_k(t_i)}{t_{i+1}-t_i}(t-t_i), \quad t\in[t_i,t_{i+1}]. \tag{1.39}$$

Il est alors possible de définir l'estimateur d'Horvitz–Thompson de la courbe moyenne basé sur des observations discrétisées

$$\widehat{\mu}_d(t) = \frac{1}{N}\sum_{k\in s}\frac{\tilde{Y}_k(t)}{\pi_k}, \quad t\in[0,T]. \tag{1.40}$$

Afin d'obtenir les propriétés de consistance de nos estimateurs, on effectue les hypothèses supplémentaires suivantes.

(**A4**) Pour tout $k\in U$, $Y_k\in C[0,T]$, l'espace des fonctions continues sur $[0,T]$, et $\lim_{N\to\infty}\mu_N=\mu$ dans $C[0,T]$.

(**A5**) Il existe deux constantes positives C_2 et C_3 et $\beta>1/2$ tels que, pour tout N, $N^{-1}\sum_{k\in U}(Y_k(0))^2<C_2$ et $N^{-1}\sum_{k\in U}(Y_k(t)-Y_k(s))^2\le C_3|t-s|^{2\beta}$ pour tout $(s,t)\in[0,T]\times[0,T]$.

Proposition 1.2.1. *Supposons que les (A1),(A2),(A4), et (A5) soient satisfaites. Si le schéma de discrétisation vérifie* $\lim_{N\to\infty}\max_{\{i=1,\ldots,d_N-1\}}|t_{i+1}-t_i|^{2\beta}=o(n^{-1})$, *alors pour une certaine constante* C

$$\sqrt{n}\ E\left\{\sup_{t\in[0,T]}|\widehat{\mu}_d(t)-\mu_N(t)|\right\}<C.$$

La proposition 1.2.1 affirme que pour une grille suffisamment fine la vitesse de convergence paramétrique peut être obtenue de façon uniforme. La condition $\beta>1/2$ dans l'hypothèse (A5), qui exige que les trajectoires doivent être assez régulières (sans nécessairement être dérivables), n'est pas très forte puisque le cas $\beta=1/2$ correspond à des trajectoires browniennes.

On souhaite également obtenir la consistance de l'estimateur $\widehat{\gamma}_d(t,t)$ de la fonction de variance $\gamma_N(t,t)$. Pour cela, on utilise l'hypothèse (A3) sur les probabilités d'inclusion d'ordre supérieures, et on introduit des hypothèses sur les moments d'ordre 4 des trajectoires.

(**A6**) Il existe deux constantes positives C_4 et C_5, telles que
$N^{-1}\sum_{k\in U_N} Y_k(0)^4 < C_4$, et $N^{-1}\sum_{k\in U_N}\{Y_k(t)-Y_k(s)\}^4 < C_5\,|t-s|^{4\beta}$, pour tout $(s,t)\in[0,T]\times[0,T]$.

Proposition 1.2.2. *Supposons que les Hypothèses (A1),(A2) et (A4)-(A6) soient vérifiées. Si le schéma de discrétisation vérifie* $\lim_{N\to\infty}\max_{\{i=1,\ldots,d_N-1\}}|t_{i+1}-t_i| = o(1)$, *alors*

$$n\ E\left\{\sup_{t\in[0,T]}|\widehat{\gamma}_d(t,t)-\gamma_N(t,t)|\right\}\to 0,\quad N\to\infty.$$

Après avoir construit un estimateur de courbe moyenne, il est intéressant de connaître la précision de cet estimateur en construisant des bandes de confiance. De façon analogue au cas univarié où on peut obtenir un intervalle de confiance d'une moyenne estimée, on souhaite construire une bande de confiance dans laquelle se trouve la vraie courbe moyenne. Pour un risque d'erreur α fixé, on veut obtenir une bande $B(t)$ telle que

$$\mathbb{P}\left(\forall t\in[0,T]\ |\mu(t)-\widehat{\mu}(t)|\leq B(t)\right)\approx 1-\alpha. \tag{1.41}$$

La loi limite de notre estimateur de la moyenne est obtenue grâce à un Théorème Central Limite fonctionnel. Pour cela, on introduit une nouvelle hypothèse qui concerne essentiellement le plan de sondage.

(**A7**) Il existe un certain $\delta>0$, tel que $N^{-1}\sum_{k\in U_N}|Y_k(t)|^{2+\delta}<\infty$ pour tout $t\in[0,T]$, et $\{\gamma_N(t,t)\}^{-1/2}\{\widehat{\mu}_N(t)-\mu_N(t)\}\to N(0,1)$ en distribution quand N tend vers l'infini.

Proposition 1.2.3. *Si les Hypothèses (A1),(A2), et (A4)-(A7) sont vérifiées et si on suppose que les points de discrétisations vérifient* $\lim_{N\to\infty}\max_{\{i=1,\ldots,d_N-1\}}|t_{i+1}-t_i|^{2\beta} = o(n^{-1})$. *Alors*

$$\sqrt{n}\ (\widehat{\mu}_d-\mu_N)\to X \quad \textit{en distribution dans } C[0,T]$$

où X est un processus Gaussien prenant ses valeurs dans $C[0,T]$ de moyenne 0 et de fonction de covariance $\breve{\gamma}(s,t)=\lim_{N\to\infty} n\gamma_N(s,t)$.

Le début de la preuve repose sur la méthode de Cramer–Wold qui permet d'avoir accès à la normalité multivariée lorsque l'on considère des trajectoires discrètes. Des critères de tension sont ensuite utilisés (et présentées dans l'annexe A.1) dans le but d'obtenir une version fonctionnelle, et non ponctuelle, du Théorème Central Limite.

En reprenant des arguments similaires à ceux de Degras (2009), on peut alors construire des bandes de confiance asymptotiques dans le but d'évaluer la précision globale de notre estimateur. Grâce à un résultat de Landau & Shepp (1970) sur les suprema de processus Gaussiens, on déduit que pour un risque d'erreur $\alpha>0$,

$$\mathbb{P}\left[|\widehat{\mu}_d(t)-\mu_N(t)|<\{2\log(2/\alpha)\widehat{\gamma}_d(t,t)\}^{1/2},t\in[0,T]\right]\ \simeq\ 1-\alpha. \tag{1.42}$$

1.2.2 Estimateurs d'Horvitz-Thompson sur données fonctionnelles bruitées

Les résultats précédents ont été étendus au cas de courbes bruitées en proposant un lissage par polynômes locaux, dont la fenêtre est choisie par validation croisée en tenant compte des poids de sondage. De plus, la construction des bandes de confiance, obtenues *via* simulations d'un processus Gaussien conditionnellement à une estimation de la fonction de covariance, atteignent le taux de couverture nominale. Cela est développé dans le chapitre 4.

L'hypothèse d'absence d'erreur de mesure n'est pas forcément vérifiée et on peut être confronté à des données bruitées. Pour toutes les unités $k \in s$, on observe

$$Y_{jk} = X_k(t_j) + \epsilon_{jk} \tag{1.43}$$

où les erreurs de mesure ϵ_{jk} sont des variables aléatoires centrées qui sont indépendantes entre les unités k mais pas nécessairement selon j. On autorise donc une dépendance temporelle du bruit. On suppose également que l'échantillon aléatoire s est indépendant du bruit ϵ_{jk} et que les trajectoires $X_k(t), t \in [0,T]$ sont déterministes.

Pour chaque unité $k \in U$, on peut alors approximer la courbe originale X_k en lissant la trajectoire discrétisée correspondante $(Y_{1k}, \ldots, Y_{dk})$ grâce à un lisseur (par exemple des splines, des noyaux ou des polynômes locaux)

$$\widehat{X}_k(t) = \sum_{j=1}^{d} W_j(t) Y_{jk}. \tag{1.44}$$

Les fonctions de poids $W_j(t)$ s'expriment, dans le cas des polynômes locaux, sous la forme

$$W_j(t) = \frac{\frac{1}{dh}\left\{s_2(t) - (t_j - t)s_1(t)\right\} K\left(\frac{t_j - t}{h}\right)}{s_2(t)s_0(t) - s_1^2(t)}, \quad j = 1, \ldots, d, \tag{1.45}$$

où K est un noyau, $h > 0$ est une fenêtre de lissage, et

$$s_l(x) = \frac{1}{dh} \sum_{j=1}^{d} (t_j - t)^l K\left(\frac{t_j - t}{h}\right), \quad l = 0, 1, 2. \tag{1.46}$$

On suppose que le noyau K est non-négatif, à support compact, et satisfait $K(0) > 0$ et $|K(s) - K(t)| \leq C|s - t|$ pour une certaine constante finie C et pour tout $s, t \in [0,T]$.

On peut alors définir l'estimateur d'Horvitz-Thompson de la courbe moyenne

$$\widehat{\mu}_N(t) \quad = \quad \frac{1}{N} \sum_{k \in s} \frac{\widehat{X}_k(t)}{\pi_k}.$$

La fonction de covariance de $\widehat{\mu}_N$ peut s'écrire

$$\mathrm{Cov}\left(\widehat{\mu}_N(s), \widehat{\mu}_N(t)\right) = \frac{1}{N} \gamma_N(s,t) \tag{1.47}$$

pour tout $s, t \in [0,T]$, où

$$\gamma_N(s,t) = \frac{1}{N} \sum_{k,l \in U} \Delta_{kl} \frac{\tilde{X}_k(s)}{\pi_k} \frac{\tilde{X}_l(t)}{\pi_l} + \frac{1}{N} \sum_{k \in U} \frac{1}{\pi_k} E\left(\tilde{\epsilon}_k(s) \tilde{\epsilon}_k(t)\right) \tag{1.48}$$

avec

$$\left\{\begin{array}{ll} \tilde{X}_k(t) & = \sum_{j=1}^d W_j(t) X_k(t_j), \\ \tilde{\epsilon}_k(t) & = \sum_{j=1}^d W_j(t) \epsilon_{kj}, \\ \Delta_{kl} & = \mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k \pi_l. \end{array}\right. \tag{1.49}$$

On peut estimer $\gamma_N(s,t)$ par

$$\widehat{\gamma}_N(s,t) = \frac{1}{N} \sum_{k,l \in U} \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{I_k}{\pi_k} \frac{I_l}{\pi_l} \right) \widehat{X}_k(s) \widehat{X}_l(t). \tag{1.50}$$

Les estimateurs de la moyenne $\widehat{\mu}_N$ et de sa fonction de covariance $\widehat{\gamma}_N$ sont asymptotiquement sans biais et consistants.

Lorsqu'on lisse les trajectoires avec un paramètre de lissage h, la qualité de l'estimateur dépend beaucoup du choix de h et un critère de sélection est donc nécessaire. Quand $\sum_{k \in s} \pi_k^{-1} = N$, comme c'est le cas pour le sondage aléatoire simple sans remise ou le sondage stratifié, on peut facilement vérifier que $\widehat{\mu}_s = \sum_{k \in s} X_k / \pi_k$ est l'argument minimum des moindres carrés fonctionnels pondérés

$$\sum_{k \in s} w_k \int_0^T \left(X_k(t) - \mu(t) \right)^2 dt \tag{1.51}$$

en supposant que $\mu \in L^2([0,T])$, et où les poids sont $w_k = (N\pi_k)^{-1}$. Alors, une méthode simple et naturelle pour sélectionner la fenêtre h est de considérer le critère de validation croisée suivant

$$\mathrm{WCV}(h) = \sum_{k \in s} w_k \sum_{j=1}^d \left(Y_{jk} - \widehat{\mu}_N^{-k}(t_j) \right)^2. \tag{1.52}$$

où

$$\widehat{\mu}_N^{-k}(t) = \sum_{\ell \in s, \ell \neq k} \widetilde{w}_\ell \ \widehat{X}_\ell(t),$$

avec des nouveaux poids $\widetilde{w}_\ell$. Cette méthode de validation croisée pondérée est plus simple que la validation croisée basée sur l'estimation de la variance proposée par Opsomer & Miller (2005). En effet, dans leur cas, le biais peut être non négligeable et s'intéresser uniquement à la partie variance de l'erreur conduit à sélectionner des valeurs de lissage trop grandes. Par ailleurs, Opsomer & Miller (2005) suggèrent de considérer les poids définis par $\widetilde{w}_\ell = w_\ell / (1 - w_k)$. Pour le sondage aléatoire simple sans remise, puisque $w_k = n^{-1}$ on a $\widetilde{w}_k = (n-1)^{-1}$, et on retrouve donc le critère de validation croisée défini en (1.52) qui est exactement le critère de validation croisée introduit par Rice & Silverman (1991) dans le cas indépendant.

Pour le sondage stratifié, une meilleure approximation qui garde la propriété de *design-based* de l'estimateur $\widehat{\mu}_N^{-k}$ peut être obtenue en tenant compte du taux de sondage dans les différentes strates. Si on a H strates de tailles N_h, $h = 1, \ldots, H$ et on tire n_h observations, *via* un sondage aléatoire simple sans remise, dans chaque strate h. Si une unité k vient de la strate h, on a $w_k = N_h (N n_h)^{-1}$. Ainsi, on prend $\widetilde{w}_\ell = (N_h - 1)\{(N-1)(n_h - 1)\}^{-1}$ pour toutes les unités $\ell \neq k$ dans la strate h et on ajuste les poids pour toutes les unités ℓ' de l'échantillon qui ne font pas partie de la strate h, $\widetilde{w}_{\ell'} = N(N-1)^{-1} w_{\ell'}$.

On cherche également à construire des bandes de confiance pour μ_N de la forme

$$\left\{ \left[\widehat{\mu}_N(t) \pm c \, \frac{\widehat{\sigma}_N(t)}{N^{1/2}} \right], \, t \in [0,T] \right\}, \tag{1.53}$$

où c un nombre à déterminer et $\widehat{\sigma}_N(t) = \widehat{\gamma}_N(t,t)^{1/2}$. Plus précisément, étant donné un niveau de confiance $1-\alpha \in [0,1]$, on cherche $c = c_\alpha$ qui vérifie approximativement

$$\mathbb{P}\left(|G(t)| \leq c\,\sigma(t), \, \forall t \in [0,T]\right) = 1-\alpha, \tag{1.54}$$

où G est une processus Gaussien de moyenne zéro et de fonction de covariance γ, et où $\sigma(t) = \gamma(t,t)^{1/2}$. Calculer de telles bandes de façon précise et explicite dans un contexte général est un problème difficile qui demanderait de fortes conditions de stationnarité qui n'ont aucune raison d'être satisfaites dans notre cas.

Néanmoins, on propose une méthode d'estimation de c par simulation, qui repose sur la normalité asymptotique de notre estimateur conditionnellement à sa fonction de covariance. On se place de nouveau dans le cadre du modèle de superpopulation et en reprenant les hypothèses effectuées sur le plan de sondage. On suppose de plus que :

- (**A8**) (*Trajectoires*) Il existe des constantes C_6 et C_7 telles que $|X_k(s) - X_k(t)| \leq C_6|s-t|^\beta$ et $|X_k(0)| \leq C_7$ pour tout $k \in U_N$, $N \geq 1$, et $s,t \in [0,T]$, où $\beta > \frac{1}{2}$ est une constante finie.
- (**A9**) (*Discrétisation*) Il existe des constantes C_8 et C_9 telles que $C_8 \leq d(t_{j+1} - t_j) \leq C_9$ pour tout $1 \leq j \leq d$, $N \geq 1$, et $\dfrac{d(\log\log N)}{N} \to 0$ quand $N \to \infty$.
- (**A10**) (*Erreurs de mesure*) Les vecteurs aléatoires $(\epsilon_{k1}, \ldots, \epsilon_{kd})', k \in U_N$, sont i.i.d. et sont distribués selon un vecteur Gaussien de moyenne zéro et de matrice de covariance $\mathbf{V}_N$. La plus grande valeur propre de la matrice de covariance vérifie $\|\mathbf{V}_N\| \leq C$ pour tout $N \geq 1$.
- (**A11**) (*TCL univarié*) Pour chaque $t \in [0,T]$ fixé, on a que

$$\frac{\widehat{\mu}_N(t) - \mu_N(t)}{\sqrt{\mathrm{Var}(\widehat{\mu}_N(t))}} \rightsquigarrow N(0,1)$$

  quand $N \to \infty$, où $\rightsquigarrow$ désigne la convergence en distribution.

Théorème 1.2.1. *On suppose que les conditions (A1),(A2), et (A8)-(A11) sont satisfaites et $dh^{1+\alpha} \to \infty$ pour un certain $\alpha > 0$ quand $N \to \infty$. Soit G un processus Gaussien de moyenne zéro et de fonction de covariance γ. Soit $(\widehat{G}_N)$ une suite de processus tel que pour chaque N, conditionnellement à $\widehat{\gamma}_N$, $\widehat{G}_N$ est processus Gaussien de moyenne zéro et de covariance $\widehat{\gamma}_N$ définie dans* (1.50)*. Alors pour tout $c > 0$, quand $N \to \infty$, :*

$$\mathbb{P}\left(|\widehat{G}_N(t)| \leq c\,\widehat{\sigma}_N(t), \, \forall t \in [0,T] \,|\, \widehat{\gamma}_N\right) \to \mathbb{P}\left(|G(t)| \leq c\,\sigma(t), \, \forall t \in [0,T]\right) \textit{ en probabilités.}$$

L'importance pratique de ce théorème est qu'il permet d'estimer le seuil c introduit en (1.54) par simulation. Conditionnellement à $\widehat{\gamma}_N$, on peut simuler un grand nombre d'échantillons de processus Gaussiens $(\widehat{G}_N/\widehat{\sigma}_N)$ et d'estimer leur norme sup. On obtient ainsi une bonne approximation de la distribution de $\|\widehat{G}_N/\widehat{\sigma}_N\|_\infty$, et il ne reste plus qu'à choisir c comme le quantile d'ordre $(1-\alpha)$ de la distribution suivante :

$$\mathbb{P}\left(|\widehat{G}_N(t)| \leq c\,\widehat{\sigma}_N(t), \, \forall t \in [0,T] \,|\, \widehat{\gamma}_N\right) = 1-\alpha. \tag{1.55}$$

1.2.3 Prise en compte de variable auxiliaire

Afin d'améliorer la précision de nos estimateurs, la prise en compte de variables auxiliaires a été étudiée pour construire un estimateur de la courbe moyenne de type *model assisted* et a été comparée avec l'estimateur stratifié dans le chapitre 5. On propose de tenir compte d'information auxiliaire réelle ou multivariée disponible sur l'ensemble de la population, avec une approche modèle assisté semi-paramétrique, afin d'améliorer la précision des estimateurs de Horvitz-Thompson de la courbe moyenne. Une possibilité est d'estimer d'abord les composantes principales fonctionnelles avec un point de vue *design-based* dans le but de réduire la dimension des courbes et d'ensuite proposer des modèles semi-paramétriques sur ces composantes principales afin d'obtenir des estimations des courbes qui ne font pas partie de l'échantillon.

Comme dans Cardot *et al.* (2010) on voudrait décrire les variations individuelles autour de la fonction moyenne dans un espace de fonctions dont la dimension est aussi petite que possible au sens d'une erreur quadratique. Considérons un ensemble de q fonctions orthornormales de $L^2[0,T]$, $\phi_1, \ldots, \phi_q$, qui minimisent le reste $R(q)$ de la projection de Y_k sur l'espace engendré par ces q fonctions

$$R(q) = \frac{1}{N} \sum_{k \in U} \|R_{qk}\|^2$$

avec $R_{qk}(t) = Y_k(t) - \mu(t) - \sum_{j=1}^{q} \langle Y_k - \mu, \phi_j \rangle \phi_j(t),\ t \in [0,T]$. Il est facile de montrer (Cardot *et al.*, 2010) que $R(q)$ atteint son minimum lorsque $\phi_1, \ldots, \phi_q$ sont les fonctions propres de l'opérateur de covariance Γ associé aux plus grandes valeurs propres, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q \geq 0$,

$$\Gamma\phi_j(t) = \int_0^T \gamma(s,t)\phi_j(s)ds = \lambda_j \phi_j(t), \quad t \in [0,T], j \geq 1.$$

Quand les observations sont issues d'un échantillon s, l'estimateur de la fonction de covariance

$$\widehat{\gamma}(s,t) \quad = \quad \frac{1}{N} \sum_{k \in s} \frac{1}{\pi_k} \left(Y_k(t) - \widehat{\mu}(t)\right) \left(Y_k(s) - \widehat{\mu}(s)\right), \ (s,t) \in [0,T] \times [0,T], \quad (1.56)$$

permet de directement obtenir des estimateurs des valeurs propres $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_q$ et des fonctions propres correspondantes $\widehat{\phi}_1, \ldots, \widehat{\phi}_q$.

Supposons à présent qu'il est possible d'avoir accès à m variables auxiliaires $X_1, \ldots, X_m$ que l'on suppose liées aux courbes individuelles Y_k et disponibles pour tous les individus k de la population. Prendre en compte cette information auxiliaire permettra certainement d'améliorer l'estimateur basique $\widehat{\mu}$. En regardant la décomposition des trajectoires individuelles Y_k sur les fonctions propres,

$$Y_k(t) = \mu(t) + \sum_{j=1}^{q} \langle Y_k - \mu, \phi_j \rangle \phi_j(t) + R_{qk}(t), \quad t \in [0,T],$$

et en reprenant des idées de Chiou *et al.* (2003) et Müller & Yao (2008), une approche intéressante consiste à modéliser les valeurs des composantes principales de la population $\langle Y_k - \mu, \phi_j \rangle$ avec les variables auxiliaires à chaque niveau j de la décomposition en fonctions

propres, $\langle Y_k - \mu, \phi_j \rangle \approx f_j(x_{k1}, \ldots, x_{km})$ où les fonctions de régression f_j peuvent être paramétriques ou non et $(x_{k1}, \ldots, x_{km})$ est le vecteur des observations des m variables auxiliaires de l'individu k.

Il est alors possible d'estimer les valeurs des composantes principales

$$\widehat{C}_{kj} = \langle Y_k - \widehat{\mu}, \widehat{\phi}_j \rangle,$$

pour $j = 1, \ldots, q$ et tout $k \in s$. Alors, un estimateur *design-based* des moindres carrés pour les fonctions f_j

$$\widehat{f}_j \quad = \quad \arg\min_{g_j} \sum_{k \in s} \frac{1}{\pi_k} \left(\widehat{C}_{kj} - g_j(x_{k1}, \ldots, x_{km}) \right)^2, \tag{1.57}$$

est pratique pour construire l'estimateur par modèle assisté $\widehat{\mu}_X$ de μ suivant

$$\widehat{\mu}_x(t) \quad = \quad \widehat{\mu}(t) - \frac{1}{N} \left(\sum_{k \in s} \frac{\widehat{Y}_k(t)}{\pi_k} - \sum_{k \in U} \widehat{Y}_k(t) \right) \tag{1.58}$$

où les courbes prédites $\widehat{Y}_k$ sont estimées pour tous les individus de la population U grâce aux m variables auxiliaires,

$$\widehat{Y}_k(t) = \widehat{\mu}(t) + \sum_{j=1}^{q} \widehat{f}_j(x_{k1}, \ldots, x_{km}) \; \widehat{v}_j(t), \; t \in [0, T].$$

Dans le chapitre 6, une comparaison est effectuée entre l'estimateur à probabilités inégales et un autre estimateur de type *model assisted*. Une seconde comparaison porte sur les bandes de confiance construites par l'estimation de la borne supérieure de processus Gaussiens et par bootstrap.

L'estimateur à probabilités inégales proportionnelles à la taille, présenté dans la section 1.1.3, se généralise directement au cas fonctionnel puisqu'il s'agît d'un estimateur de Horvitz-Thompson avec des poids π_k déterminés par avance. Il est possible d'estimer sa fonction de covariance grâce à la formule d'Hájek (1.24) étendue par analogie au cas fonctionnel. Au lieu d'utiliser l'information auxiliaire directement dans le plan de sondage, on peut également ajuster un modèle linéaire et construire un autre estimateur par modèle assisté $\widehat{\mu}_{\mathrm{ma}}$. Plus précisément, on peut écrire pour toutes les unités k et $t \in [0, T]$

$$Y_k(t) = \beta_0(t) + \beta_1(t) x_k + \epsilon_{kt} \tag{1.59}$$

où $\beta_0(t)$ et $\beta_1(t)$ sont des coefficients de régressions (voir Faraway, 1997). Les poids du plan de sondage peuvent être pris en compte pour estimer $\widehat{\beta}_0$ et $\widehat{\beta}_1$ de β_0 et β_1 (voir Särndal *et al.*, 1992). Finalement, on obtient l'estimateur de la courbe moyenne pour tout $t \in [0, T]$,

$$\widehat{\mu}_{\mathrm{ma}}(t) = \frac{1}{N} \sum_{k \in s} \frac{Y_k(t)}{\pi_k} - \frac{1}{N} \left(\sum_{k \in s} \frac{\widehat{Y}_k(t)}{\pi_k} - \sum_{k \in U} \widehat{Y}_k(t) \right) \tag{1.60}$$

où $\widehat{Y}_k(t) = \widehat{\beta}_0(t) + \widehat{\beta}_1(t) x_k, \; t \in [0, T]$. La covariance asymptotique de cet estimateur peut être estimée par une formule analogue à celle de Breidt & Opsomer (2000).

Bibliographie

Baíllo, A. Cuevas, A. and Fraiman, R. (2010). Classification méthods for functional data. *The Oxford Handbook of Functional Data Analysis*, Edited by Ferraty, F. and Romain, Y., **10**, 259-297.

Besse, P.C. Cardot, H. and D. Stephenson (2000). Autoregressive Forecasting of Some Functional Climatic Variations. *Scandinavian Journal of Statistics*, Vol. **27**, 673-687.

Bickel, P. J. & Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics*, **12**, 470-482.

Breidt, F. J. & Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, **28**, 1026-1053.

Bunea, F., Ivanescu, A. and M. Wegkamp (2011). Adaptive inference for the mean of a Gaussian process in functional data. *J. Roy. Statist. Soc. Ser. B*, to appear.

Cardot, H., Ferraty, F. and P. Sarda (1999). Functional Linear Model. *Statistics and Probability Letters*. Vol. 45, **1**, 11-22.

Cardot, H., Faivre, R. and M. Goulard (2003). Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data. *Journal of Applied Statistics*, Vol. **30**, 1185-1199.

Cardot, H., Ferraty, F. and P. Sarda (2003). Spline Estimators for the Functional Linear Model. *Statistica Sinica*, Vol. 13, 571-591.

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.

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

Chauvet, G. (2007). Méthodes de Bootstrap en population finie. *PhD Thesis*, Université Rennes II, France.

Chiky, R. (2009). Résumé de flux de données distribuées. *PhD Thesis*, l'École Nationale Supérieure des Télécommunications, France.

Chiou, J.M., Müller, H.G. and Wang, J.L. (2003). Functional quasi-likelihood regression models with smooth random effects. *J.Roy. Statist. Soc., Ser. B*, **65**, 405-423.

Cochran, W.G. (1977). *Sampling techniques.* 3rd Edition, Wiley, New York.

Dauxois J. & Pousse A. (1976). Les analyses factorielles en calcul des probabilités et en statistiques : essai d'étude synthétique. *Thèse d'état*, Université Paul Sabatier, Toulouse, France.

Degras, D. (2009). Nonparametric estimation of a trend based upon sampled continuous processes. *C. R. Math. Acad. Sci. Paris*, **347**, 191-194.

Deville J.C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE*, **15**, 3-101.

Deville, J.C. & Tillé, Y. (2000). Balanced sampling by means of the cube method. Document de travail, Rennes, CREST-ENSAI.

Efron, B. (1979). Bootstrap methods : another look at the jackknife. *Annals of Statistics*, **7** ,1–26.

Erdös, P. & Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hungar. Acad. Sci.* **4**, 49-61.

Ferraty, F. & Vieu, P. (2006). Nonparametric functional data analysis : theory and practice. *Springer Series in Statitics*, Springer.

Ferraty, F. & Vieu, P. (2010). Kernel regression estimation for functional data. *The Oxford Handbook of Functional Data Analysis*, Edited by Ferraty, F. and Romain, Y., **4**, 72-129.

Fuller, W.A. (2009). *Sampling Statistics.* John Wiley and Sons.

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

Hàjek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hungar. Acad. Sci.* **5**, 361-374.

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

Hall, P. (2010). Principal component analysis for functional data. *The Oxford Handbook of Functional Data Analysis*, Edited by Ferraty, F. and Romain, Y., **8**, 210-234.

Hall, P., Müller, H.G. and Wang, J.L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.*, Vol 34, Number 3, 1493-1517.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Isaki, C.T. & Fuller, W.A. (1982). Survey design under the regression superpopulation model. *J. Am. Statist. Ass.* **77**, 89-96.

James, G. (2010). Sparseness and functional data analysis. *The Oxford Handbook of Functional Data Analysis*, Edited by Ferraty, F. and Romain, Y., **11**, 298-323.

Kneip, A. and Utikal, K. J. (2001). Inference for Density Families Using Functional Principal Component Analysis. *Journal of the American Statistical Association*, **96**, 519–542.

Krewski, D. and Rao, J. (1981). Inference from stratified samples : properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, **9**, 1010–1019.

Landau, H. & Shepp, L.A. (1970). On the supremum of a Gaussian process. *Sankhyã*, **32**, 369-378.

Mas A. (2007). Testing for the mean of random curves : a penalization approach. *Statistical Inference for Stochastic Processes,* **10**, 147-163

Müller, H.G. & Stadtmüller, U. (2005) Generalized functional linear models. *Ann. Statist.*, **33**, 774-805.

Müller, H.G. Leng, X. (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, **22**, 68-76.

Müller, H.G., Yao, F. (2008). Functional Additive Model. *J. Am. Statist. Ass.* **103**, 1534-1544.

Neyman, J. (1934). On the two different aspects of representative method : the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, **97**, 558-606.

Opsomer, J. D. and Miller, C. P. (2005). Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. *J. Nonparametric Statistics,* **17**, 593-611.

Preda, C. & Saporta, G. (2002). Régression PLS sur une processus stochastique. *Revue de statistique appliquée*, **50**, 27-45.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B*, **53**, 233-243.

Rice, J. and Wu, C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253-259.

Robinson, P. M. & Särndal, C. E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya : The Indian Journal of Statistics*, **45**, 240-248.

Särndal, C. E., Swensson, B. and J. Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag.

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

Sood, A., James, G. and Tellis, G. (2009). Functional Regression : A New Model for Predicting Market Penetration of New Products. *Marketing Science*, **28**, 36-51.

Tillé, Y. (2001). Théorie des sondages : Échantillonnage et estimation en populations finies. *Dunod*, Paris.

Tillé, Y. (2006). Sampling Algorithms. *Springer*, New York.

Yates, F. and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, *B15*, 235–261.

Yao, F. Müeller, H.G. and Wang, J.L. (2005). Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, 577-590.

Chapitre 2

Horvitz–Thompson estimators for functional data : asymptotic confidence bands and optimal allocation for stratified sampling *

Résumé : Lorsque l'on travaille sur de grandes bases de données de données fonctionnelles, les techniques de sondage sont utiles pour obtenir des estimateurs de quantités fonctionnelles simples, sans être obligé de stocker toutes les données. Nous proposons ici un estimateur d'Horvitz–Thompson de la trajectoire moyenne. Dans le cadre du modèle de superpopulation, nous montrons sous des conditions de régularité faibles que nous obtenons des estimateurs de la fonction moyenne et de leurs fonctions de variance, qui sont uniformément consistants. Grâce à des hypothèses supplémentaires sur le plan de sondage, nous établissons un Théorème Central Limite Fonctionnel et déduisons des bandes de confiance asymptotiques. Le sondage stratifié est étudié en détails, et nous obtenons également une version fonctionnelle de la règle d'allocation optimale habituelle en considérant un critère de variance moyenne. Ces techniques sont illustrées au moyen d'une population test de $N = 18902$ compteurs électriques pour lesquels nous avons la consommation électrique individuelle mesurée toutes les 30 minutes pendant une semaine. Nous remarquons que la stratification peut substantiellement améliorer à la fois la précision des estimateurs et réduire la largeur de la bande de confiance comparée au sondage aléatoire simple sans remise.

Mots clés : Variance asymptotique, Théorème Central Limite Fonctionnel, Modèle de superpopulation, Suprema de processus Gaussiens, Échantillonnage.

∗. Article écrit en collaboration avec Hervé Cardot et publié dans le journal *Biometrika* sous la référence suivante :
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.
Le chapitre 3 contient quelques détails techniques complémentaires qui n'ont pas été publiés.

Horvitz–Thompson estimators for functional data: asymptotic confidence bands and optimal allocation for stratified sampling

By HERVÉ CARDOT and ETIENNE JOSSERAND
Institut de Mathématiques de Bourgogne, UMR CNRS 5584,
Université de Bourgogne
9 Avenue Alain Savary - B.P. 47870, 21078 DIJON Cedex - France
herve.cardot@u-bourgogne.fr etienne.josserand@u-bourgogne.fr

Summary

When dealing with very large datasets of functional data, survey sampling approaches are useful in order to obtain estimators of simple functional quantities, without being obliged to store all the data. We propose here a Horvitz–Thompson estimator of the mean trajectory. In the context of a superpopulation framework, we prove under mild regularity conditions that we obtain uniformly consistent estimators of the mean function and of its variance function. With additional assumptions on the sampling design we state a functional Central Limit Theorem and deduce asymptotic confidence bands. Stratified sampling is studied in detail, and we also obtain a functional version of the usual optimal allocation rule considering a mean variance criterion. These techniques are illustrated by means of a test population of $N = 18902$ electricity meters for which we have individual electricity consumption measures every 30 minutes over one week. We show that stratification can substantially improve both the accuracy of the estimators and reduce the width of the global confidence bands compared to simple random sampling without replacement.

Some key words: Asymptotic variance; Functional Central Limit Theorem; Superpopulation model; Supremum of Gaussian processes; Survey sampling.

2.1 Introduction

The development of distributed sensors has enabled access to potentially huge databases of signals evolving along time and observed on very fine scales. Exhaustive collection of such data would require major investments, both for transmission of the signals through networks and for storage. As noted in Chiky & Hébrail (2009), survey sampling of the sensors, which entails randomly selecting only a part of the curves of the population and which represents a trade off between limited storage capacities and the accuracy of the data, may be relevant compared to signal compression in order to obtain accurate approximations to simple functional quantities such as mean trajectories.

Our study is motivated by the estimation, in a fixed time interval, of the mean electricity consumption curve of a large number of consumers. The French electricity operator EDF, Électricité De France, intends over the next few years to install over 30 million electricity meters, in each firm and household, which will be able to send individual electricity consumption measures on very fine time scales. Collecting, saving and analyzing all

this information, which may be considered as functional, would be very expensive. As an illustrative example, a sample of 20 individual curves, selected among a test population of $N = 18902$ electricity meters, is plotted in Figure 2.1. The curves consist, for each company selected, of the electricity consumption measured every 30 minutes over a period of one week. The target is the mean population curve, and we note the high variability between individuals.

Using survey sampling strategies is one way to get accurate estimates at reasonable cost. The main questions addressed in this paper are to determine the precision of a survey sampling strategy and the strategies likely to improve the sampling selection process in order to obtain estimators that are as accurate as possible and to derive global confidence bands that are as sharp as possible for stratified sampling. There is a vast literature in survey sampling theory ; see for example Fuller (2009). However, as far as we know, the convergence issue with such sampling strategies in finite population has not yet been studied in the functional data analysis literature (Ramsay & Silverman, 2005, Müller, 2005) except by Cardot *et al.* (2010), where the objective was to reduce the dimension of the data through functional principal components in the Hilbert space of square integrable functions. Here we adopt a different point of view and consider the sampled trajectories as elements of the space of continuous functions equipped with the usual sup norm in order to get uniform consistency results through maximal inequalities. Then, it is possible to build global confidence bands with the help of properties of suprema of Gaussian processes and the functional central limit theorem.

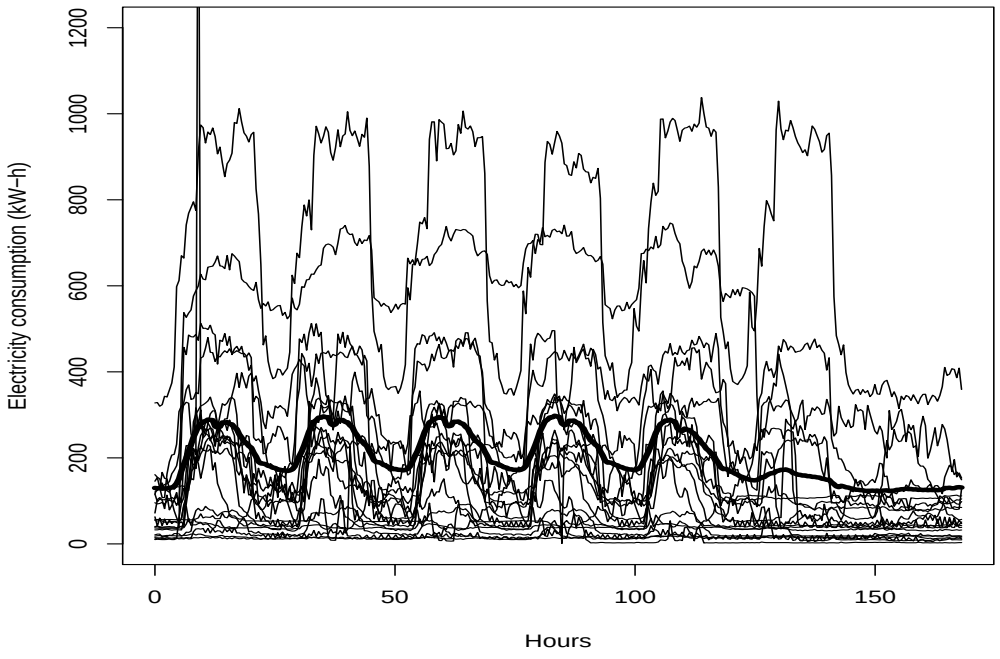

Figure 2.1: A sample of 20 individual electricity consumption curves. The mean profile is plotted in bold line.

2.2 Notation, estimators and basic properties

Let us consider a finite population $U_N = \{1, \ldots, k, \ldots, N\}$ of size N, and suppose that to each unit k in U_N we can associate a unique function $Y_k(t)$, for $t \in [0, T]$, with $T < \infty$. Our target is the mean trajectory

$$\mu_N(t) = \frac{1}{N} \sum_{k \in U} Y_k(t), \quad t \in [0, T]. \tag{2.1}$$

We consider a sample s drawn from U_N according to a fixed-size sampling design $p_N(s)$, where $p_N(s)$ is the probability of drawing the sample s. The size n_N of s is non-random and we suppose that the first and second order inclusion probabilities satisfy $\pi_k = \text{pr}(k \in s) > 0$, for all $k \in U_N$, and $\pi_{kl} = \text{pr}(k \,\&\, l \in s) > 0$ for all $k, l \in U_N$, $k \neq l$, so that each unit and each pair of units can be drawn with a non null probability from the population.

It is now possible to write the classical Horvitz–Thompson estimator of the mean curve,

$$\widehat{\mu}_N(t) = \frac{1}{N} \sum_{k \in U} \frac{Y_k(t)}{\pi_k} I_k, \quad t \in [0, T], \tag{2.2}$$

where I_k is the sample membership indicator, $I_k = 1$ if $k \in s$ and $I_k = 0$ otherwise. We clearly have $E(I_k) = \pi_k$ and $E(I_k I_l) = \pi_{kl}$.

It is easy to check (Fuller, 2009) that this estimator is unbiased, *i.e.* for all $t \in [0, T]$, $E\{\widehat{\mu}_N(t)\} = \mu_N(t)$. Its covariance function $\gamma_N(s, t) = \text{cov}\,\{\widehat{\mu}_N(s), \widehat{\mu}_N(t)\}$ satisfies, for all $(s, t) \in [0, T] \times [0, T]$,

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

with $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$ if $k \neq l$ and $\Delta_{kk} = \pi_k(1 - \pi_k)$. An unbiased estimator of $\gamma_N(s, t)$, for all $(s, t) \in [0, T] \times [0, T]$, is

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

With real data, such as the electricity consumption trajectories presented in Fig. 2.1, we do not observe $Y_k(t)$ at every instant t in $[0, T]$ but only have an evaluation of Y_k at d discretization points $0 = t_1 < \cdots < t_d = T$. Assuming that there are no measurement errors, which seems realistic in the case of electricity consumption curves, and that the trajectories are regular enough, linear interpolation is a robust and simple way to obtain accurate approximations of the trajectories at every instant t. For each unit k in the sample s, the interpolated trajectory is defined by

$$\tilde{Y}_k(t) = Y_k(t_i) + \frac{Y_k(t_{i+1}) - Y_k(t_i)}{t_{i+1} - t_i}(t - t_i), \quad t \in [t_i, t_{i+1}]. \tag{2.3}$$

It is then possible to define the Horvitz–Thompson estimator of the mean curve based on the discretized observations as

$$\widehat{\mu}_d(t) = \frac{1}{N} \sum_{k \in s} \frac{\tilde{Y}_k(t)}{\pi_k}, \quad t \in [0, T]. \tag{2.4}$$

The covariance function of $\widehat{\mu}_d$, denoted by $\gamma_d(s,t) = \text{cov}\,\{\widehat{\mu}_d(s), \widehat{\mu}_d(t)\}$, also satisfies for all $(s,t) \in [0,T] \times [0,T]$,

$$\gamma_d(s,t) = \frac{1}{N^2} \sum_{k \in U_N} \sum_{l \in U_N} \frac{\tilde{Y}_k(s)}{\pi_k} \frac{\tilde{Y}_l(t)}{\pi_l} \Delta_{kl}, \tag{2.5}$$

and, as above, an unbiased estimator of $\gamma_d(s,t)$ is

$$\widehat{\gamma}_d(s,t) = \frac{1}{N^2} \sum_{k \in s} \sum_{l \in s} \frac{\tilde{Y}_k(s)}{\pi_k} \frac{\tilde{Y}_l(t)}{\pi_l} \frac{\Delta_{kl}}{\pi_{kl}} \,. \tag{2.6}$$

To go further we must adopt an asymptotic point of view assuming that the size N of the population grows to infinity.

2.3 Asymptotic Properties

2.3.1 Assumptions

Let us consider the superpopulation asymptotic framework introduced by Isaki & Fuller (1982) and discussed in detail in Fuller (2009). We consider a sequence of growing and nested populations U_N with size N tending to infinity and a sequence of samples s_N of size n_N drawn from U_N according to the fixed-size sampling designs $p_N(s_N)$. Let us denote by π_{kN} and π_{klN} their first and second order inclusion probabilities. The sequence of subpopulations is an increasing nested one while the sample sequence is not. For simplicity of notation, we drop the subscript N in the following when there is no ambiguity. To prove our asymptotic results, we make the following assumptions.

Assumption 1. We assume that $\lim_{N \to \infty} \frac{n}{N} = \pi \in]0,1[$.

Assumption 2. We assume that $\min_k \pi_k \geq \lambda > 0$, $\min_{k \neq l} \pi_{kl} \geq \lambda * > 0$,

$$\limsup_{N \to \infty} \; n \max_{k \neq l} |\pi_{kl} - \pi_k \pi_l| < C_1 < \infty.$$

Assumption 3. For all $k \in U$, $Y_k \in C[0,T]$, the space of continuous functions on $[0,T]$, and $\lim_{N \to \infty} \mu_N = \mu$ in $C[0,T]$.

Assumption 4. There are two positive constants C_2 and C_3 and $\beta > 1/2$ such that, for all N, $N^{-1} \sum_{k \in U} (Y_k(0))^2 < C_2$ and $N^{-1} \sum_{k \in U} (Y_k(t) - Y_k(s))^2 \leq C_3 |t-s|^{2\beta}$ for all $(s,t) \in [0,T] \times [0,T]$.

Assumptions 1 and 2 concern the moment properties of the sampling designs and are fulfilled for sampling plans such as simple random sampling without replacement or stratified sampling (Robinson & Särndal, 1983, Breidt & Opsomer, 2000). Assumptions 3 and 4 are of a functional nature and seem to be rather weak. Assumption 3 imposes only that the limit of the mean function exists and is continuous, and Assumption 4 states that the trajectories have a uniformly bounded second moment and their mean squared increments satisfy a Hölder condition.

2.3.2 Consistency

We can now state the first consistency results, assuming that the grid of the d_N discretization points becomes finer and finer in $[0,T]$ as the population size N tends to infinity.

Proposition 2.3.1. *Let Assumptions 1-4 hold. If the discretization scheme satisfies* $\lim_{N\to\infty} \max_{\{i=1,\ldots,d_N-1\}} |t_{i+1}-t_i|^{2\beta} = o(n^{-1})$, *then for some constant* C

$$\sqrt{n}\ E\left\{\sup_{t\in[0,T]} |\widehat{\mu}_d(t)-\mu_N(t)|\right\} < C.$$

Proposition 2.3.1 states that if the grid is fine enough then classical parametric rates of convergence can be attained uniformly, the additional hypothesis meaning that for smoother trajectories, *i.e.* larger β, fewer discretization points are needed. We would also like to obtain that $\widehat{\gamma}_d(t,t)$ is a consistent estimator of the variance function $\gamma_N(t,t)$. To do so, we need to introduce additional assumptions concerning the higher-order inclusion probabilities and the fourth order moments of the trajectories.

Assumption 5. We assume that

$$\lim_{N\to\infty} \max_{(i_1,i_2,i_3,i_4)\in D_{4,N}} |E\{(I_{i_1}I_{i_2}-\pi_{i_1i_2})(I_{i_3}I_{i_4}-\pi_{i_3i_4})\}| = 0,$$

where $D_{t,N}$ denotes the set of all distinct t-tuples $(i_1,\ldots,i_t)$ from U_N.
We also suppose that there are two positive constants C_4 and C_5, such that $N^{-1}\sum_{k\in U_N} Y_k(0)^4 < C_4$, and $N^{-1}\sum_{k\in U_N}\{Y_k(t)-Y_k(s)\}^4 < C_5\,|t-s|^{4\beta}$, for all $(s,t)\in[0,T]\times[0,T]$.

The first part of Assumption 5 is more restrictive than Assumption 2 and is assumed, for example, in Breidt & Opsomer (2000, part of assumption (A7)). It holds, for instance, in simple random sampling without replacement and stratified sampling.

Proposition 2.3.2. *Let Assumptions 1-5 hold. If the discretization scheme satisfies* $\lim_{N\to\infty} \max_{\{i=1,\ldots,d_N-1\}} |t_{i+1}-t_i| = o(1)$, *then*

$$n\ E\left\{\sup_{t\in[0,T]} |\widehat{\gamma}_d(t,t)-\gamma_N(t,t)|\right\} \to 0, \quad N\to\infty.$$

The multiplier n that appears in the Proposition 2.3.2 is due to the fact $n\gamma_N(t,t)$ is a bounded function. Proposition 2.3.2 only states that we can obtain a uniformly consistent estimator of the variance function of the estimated mean trajectory. More restrictive conditions concerning the sampling design would be needed to get rates of convergence.

2.3.3 Asymptotic normality and confidence bands

Proceeding further, we would now like to derive the asymptotic distribution of our estimator $\widehat{\mu}_d$ in order to build asymptotic confidence intervals and bands. Obtaining the asymptotic normality of estimators in survey sampling is a technical and difficult issue

even for simple quantities such as means or totals of real numbers. Although confidence intervals are commonly used in the survey sampling community, the Central Limit Theorem has only been checked rigourously, as far as we know, for a few sampling designs. Erdös & Rényi (1959) and Hàjek (1960) proved that the Horvitz–Thompson estimator is asymptotically Gaussian for simple random sampling without replacement. These results were extended more recently to stratified sampling by Bickel & Freedman (1994) and some particular cases of two-phase sampling designs by Chen & Rao (2007). Fuller (2009, §1.3) proposes a recent review. Let us assume that the Horvitz–Thompson estimator satisfies a Central Limit Theorem for real valued quantities with new moment conditions.

Assumption 6. There is some $\delta > 0$, such that $N^{-1}\sum_{k\in U_N}|Y_k(t)|^{2+\delta} < \infty$ for all $t \in [0,T]$, and $\{\gamma_N(t,t)\}^{-1/2}\{\widehat{\mu}_N(t) - \mu_N(t)\} \to N(0,1)$ in distribution when N tends to infinity.

We can now formulate the following proposition, which tells us that if the sampling design is such that the Horvitz–Thompson estimator of the total of real quantities is asymptotically Gaussian, then our estimator $\widehat{\mu}_d$ is also asymptotically Gaussian in the space of continuous functions equipped with the sup norm. This means that point-wise normality can be transposed, under regularity assumptions on the trajectories and the asymptotic distance between adjacent discretization points, to a functional Central Limit Theorem.

Proposition 2.3.3. *Let Assumptions 1-4 and 6 hold and suppose that the discretization points satisfy* $\lim_{N\to\infty}\max_{\{i=1,\ldots,d_N-1\}}|t_{i+1}-t_i|^{2\beta} = o(n^{-1})$. *We then have that*

$$\sqrt{n}\ (\widehat{\mu}_d - \mu_N) \to X \quad \textit{in distribution in } C[0,T]$$

where X is a Gaussian random function taking values in $C[0,T]$ with mean 0 and covariance function $\breve{\gamma}(s,t) = \lim_{N\to\infty} n\gamma_N(s,t)$.

The proof, given in the Appendix, is based on the Cramer–Wold device which gives access to multivariate normality when considering discretized trajectories. Tightness arguments are then invoked in order to obtain the functional version of the Central Limit Theorem.

Using heuristic arguments similar to those of Degras (2009), we can also build asymptotic confidence bands in order to evaluate the global accuracy of our estimator. To do so, we make use of an asymptotic result from Landau & Shepp (1970), which states that the supremum of a centred Gaussian random function Z taking values in $C[0,T]$, with covariance function $\rho(s,t)$ satisfies

$$\lim_{\lambda\to\infty}\lambda^{-2}\log\mathrm{pr}\left\{\sup_{t\in[0,T]}Z(t) > \lambda\right\} \quad = \quad -\left\{2\mathrm{sup}_{t\in[0,T]}\rho(t,t)\right\}^{-1}. \tag{2.7}$$

Assuming that $\inf_t \breve{\gamma}(t,t) > 0$, it is easy to prove, with Slutsky's Lemma and Propositions 2.3.2 and 2.3.3, that the sequence of random functions $Z_n(t) = \{\widehat{\gamma}_d(t,t)\}^{-1/2}\{\widehat{\mu}_d(t) - \mu_N(t)\}$ satisfies the Central Limit Theorem in $C[0,T]$ and converges in distribution to $Z(t)$. Then, the continuous mapping theorem tells us that, for each $\lambda > 0$, $\mathrm{pr}\{\sup_t|Z_n(t)| > \lambda\}$ converges to $\mathrm{pr}\{\sup_t|Z(t)| > \lambda\}$. Applying (2.7) to Z_n, a direct computation yields that, for

a given risk $\alpha > 0$,

$$\text{pr}\left[|\widehat{\mu}_d(t) - \mu_N(t)| < \{2\log(2/\alpha)\widehat{\gamma}_d(t,t)\}^{1/2}, t \in [0,T]\right] \simeq 1-\alpha. \tag{2.8}$$

Equation (2.8) indicates that, compared to point-wise confidence intervals, global ones can be obtained simply by replacing the scaling given by the quantile of a normal centred unit variance Gaussian variable by the factor $\{2\log(2/\alpha)\}^{1/2}$. For example, if α=0·05, respectively α=0·01, then $\{2\log(2/\alpha)\}^{1/2}$= 2·716, respectively 3·255, instead of 1·960, respectively 2·576, for a point-wise confidence interval with 0·95 confidence, respectively 0·99. The result presented in equation (2.7) is asymptotic and is therefore more reliable when α is close to zero as seen in our simulation study.

2.4 Stratified sampling designs

We now consider now the particular case of stratified sampling with simple random sampling without replacement in all strata, assuming the population U is divided into a fixed number H of strata. This means that there is a partitioning of U into H sub-populations denoted by U_h, $(h = 1, \ldots, H)$. We can define the mean curve μ_h within each stratum h as $\mu_h(t) = N_h^{-1} \sum_{k \in U_h} Y_k(t),\ t \in [0,T]$, where N_h is the number of units in stratum h. The covariance function, $\gamma_h(s,t)$, within stratum h is defined by $\gamma_h(s,t) = N_h^{-1} \sum_{k \in U_h} \{Y_k(s) - \mu_h(s)\}\{Y_k(t) - \mu_h(t)\},\ (s,t) \in [0,T] \times [0,T]$.

In stratified sampling with simple random sampling without replacement in all strata, the first and second order inclusion probabilities are explicitly known, and the mean curve estimator of $\mu_N(t)$ is $\widehat{\mu}_{\text{strat}}(t) = N^{-1} \sum_{h=1}^{H} n_h^{-1} N_h \sum_{k \in s_h} Y_k(t),\ t \in [0,T]$, where s_h is a sample of size n_h, with $n_h < N_h$, obtained by simple random sampling without replacement in stratum U_h. The covariance function of $\widehat{\mu}_{\text{strat}}$, can be expressed as

$$\gamma_{\text{strat}}(s,t) = \frac{1}{N^2} \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} \widetilde{\gamma}_h(s,t), \quad (s,t) \in [0,T] \times [0,T],$$

with $(N_h - 1)\widetilde{\gamma}_h(s,t) = N_h \gamma_h(s,t)$.

For real valued quantities, optimal allocation rules, which determine the sizes n_h of the samples in all the strata, are generally defined in order to obtain an estimator whose variance is as small as possible. In our functional context, and as in the multivariate case (Cochran, 1977, §5A.2), determining an optimal allocation clearly depends on the criterion to be minimized. Indeed, one could consider many different optimization criteria which would lead to different optimal allocations rules. The width of the global confidence bands derived in equation (2.8) depend only on the standard deviation of the estimator at each instant t and minimising the width at the worst instant of time or minimizing the average width along time are natural criteria. Nevertheless, finding the solution of such optimization problems is not trivial and not investigated further in this paper. If we consider the optimal allocation based on minimising the mean variance instead of the mean standard deviation, we can then find explicit and simple solutions to

$$\min_{(n_1,\ldots,n_H)} \int_0^T \gamma_{\text{strat}}(t,t)\, dt \ \text{ subject to } \ \sum_{h=1}^{H} n_h = n \text{ and } n_h > 0,\ h = 1, \ldots, H. \tag{2.9}$$

The solution is

$$n_h^* \quad = \quad n \, \frac{N_h S_h}{\sum_{i=1}^{H} N_i S_i}, \tag{2.10}$$

with $S_h^2 = \int_0^T \widetilde{\gamma}_h(t,t)dt$, $h = 1, \ldots, H$, similar to that of the multivariate case when considering a total variance criterion (Cochran, 1977). This means that a stratum with higher variance than the others should be sampled at a higher sampling rate n_h/N_h. The gain when considering optimal allocation compared to proportional allocation, *i.e.* $n_h = nN_h/N$, can also be derived easily.

2.5 An illustration with electricity consumption

Over the next few years Électricité De France plans to install millions of sophisticated electricity meters that will be able to send, on request, electricity consumption measurements every second. Empirical studies have shown that even the simplest survey sampling strategies, such as simple random sampling without replacement, are very competitive with signal processing approaches such as wavelet expansions, when the aim is to estimate the mean consumption curve. To test and compare the different possible strategies, a test population of $N = 18902$ electricity meters has been installed in small and large companies. These electricity meters have read electricity consumption every half an hour over a period of two weeks.

We split the temporal observations and considered only the second week for estimation. The reading from first week were used to build the strata. Thus, our population of curves is a set of $N = 18902$ vectors $Y_k = \{Y_k(t_1), \ldots, Y_k(t_d)\}$ with sizes $d = 336$. Identifying each unit k of the population with its trajectory Y_k, we consider now a particular case of stratified sampling which consists in clustering the space $C[0,T]$ of all possible trajectories into a fixed number of H strata.

The strata were built by clustering the population according to the maximum level of consumption during the first week. We decided to retain $H = 4$ different clusters based on the quartiles so that all the strata have the same size. The mean trajectories during the first week in the clusters, drawn in Figure 2.2 (a), show a clear size effect. The strata have been numbered according to global mean consumption. Stratum 4, at the top of Figure 2.2 (a), corresponds to consumers with high global levels of consumption whereas stratum 1, at the bottom of Figure 2.2 (a), corresponds to consumers with low global levels of consumption.

We compared three sampling strategies, with the same sample size $n = 2000$, to estimate the mean population curve $\mu(t)$ and build confidence bands during the second week. In order to evaluate these estimators, we drew 1000 samples using the following sampling designs,

Design 1: simple random sampling estimator without replacement, which was first tested by Electricité de France;

Design 2: stratified sampling with proportional allocation, in which allocation in each stratum is defined as follows $n_h = nN_h/N$; the size of each stratum is 500;

Design 3: stratified sampling with optimal allocation according to the rule defined in (2.10). The sizes of the strata are 126 (stratum 1), 212 (stratum 2), 333 (stratum 3) and 1329 (stratum 4).

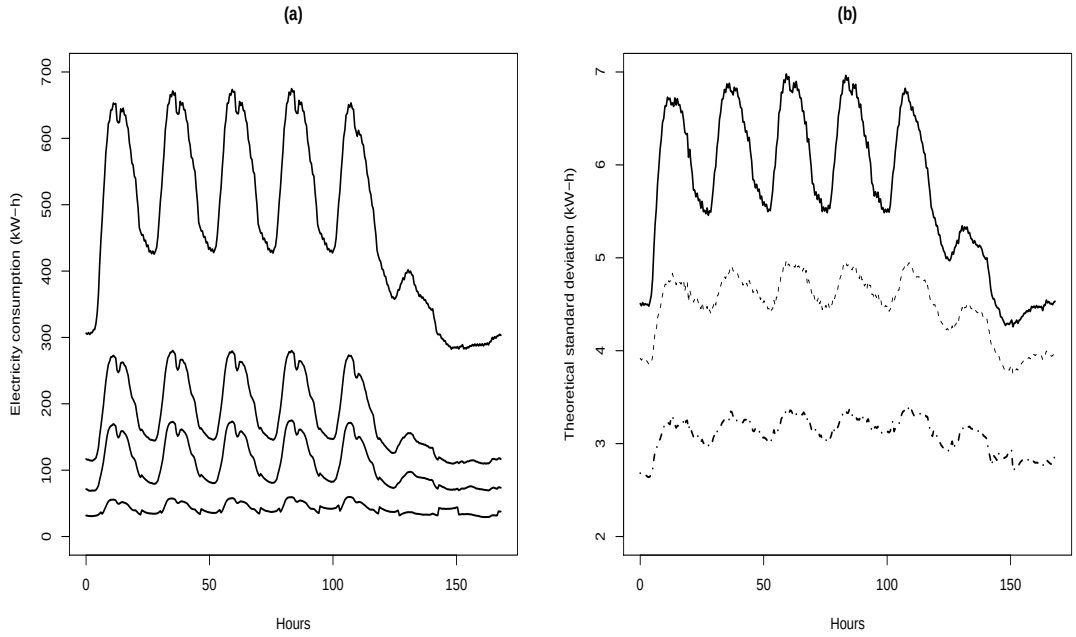

Figure 2.2: (a) Mean curve in each stratum. (b) Theoretical standard deviation function $\sqrt{\gamma(t,t)}$ for simple random sampling without replacement (solid line), stratified sampling with proportional allocation (dashed line) and stratified sampling with optimal allocation (dotted dashed line) sampling designs.

To evaluate the accuracy of the estimators, we considered the following loss criteria, evaluated with discretized data using quadrature rules, for the estimator $\widehat{\mu}$, respectively $\widehat{\gamma}$, of the mean trajectory, respectively of the mean variance,

$$R(\widehat{\mu}) \;=\; \int_0^T |\widehat{\mu}(t) - \mu(t)|\; dt, \quad R(\widehat{\gamma}) \;=\; \int_0^T |\widehat{\gamma}(t,t) - \gamma(t,t)|\; dt. \tag{2.11}$$

Basic statistics for the estimation errors of the mean function are given in Table 2.1. First, we observe that clustering the space of functions by means of stratified sampling leads to a large gain in terms of the accuracy of the estimators. In addition, there is a substantial difference between the proportional and the optimal allocation rules.

Table 2.1: Estimation errors for μ and $\gamma(t,t)$ for the different sampling designs

	Mean function				Variance function			
	Mean	1st quartile	median	3rd quartile	mean	1st quartile	median	3rd quartile
Design 1	4·46	2·37	3·75	5·68	5·26	2·42	4·04	6·64
Design 2	3·48	2·03	2·87	4·43	4·77	2·07	3·51	5·79
Design 3	2·43	1·55	2·10	3·04	1·02	0·56	0·88	1·30

We now examine the true standard deviation functions $\sqrt{\gamma(t,t)}$, which are proportional to the width of the confidence bands. They depend on the sampling design and are drawn

in Figure 2.2 (b). The theoretical standard deviation is much smaller, at all instants t, for the optimal allocation rule, and it is about twice smaller compared to simple random sampling without replacement. There is also a strong periodicity effect in the simple random sampling without replacement due to the lack of control over the units with high levels of consumption (stratum 4). Estimation errors, according to criterion (2.11), of the true covariance functions are reported in Table 2.1. The error is much smaller for stratified optimal allocation than for the other estimators; optimal allocation provides better estimates as well as better estimation of their variance.

Finally, we computed the global confidence bands to check that formula (2.8), which relies on asymptotic properties of the supremum of Gaussian processes, remains valid when considering confidence levels 0·95 and 0·99. The empirical coverage is close to the nominal one for the simple random sampling without replacement, 93·8% and 98·3%, whereas it is a little bit liberal, especially for smaller levels, for the stratified sampling designs, 88·7% and 96·8% for proportional allocation, and 88·1% and 96·8% for optimal allocation.

2.6 Concluding remarks

The experimental results on a test population of electricity consumption curves confirm that stratification, in conjunction with the optimal allocation rule, can lead, in cases of such high dimensional data, to important gains in terms of the accuracy of the estimation and width of the global confidence bands compared to more basic approaches. We have proposed a simple rule to get confidence bands that could certainly be improved, in terms of empirical coverage, by computing more realistic scaling factors with bootstrap procedures (Faraway, 1997) or Gaussian process simulations (Degras, 2010).

Choosing appropriate strata is also an important aspect of such improvement. Nevertheless, it will generally be impossible to determine for all units to which cluster they belong. Borrowing ideas from Breidt & Opsomer (2008), one possible strategy is to perform clustering on the observed sample and then try to predict to which stratum the units that are not in the sample belong using auxiliary information and supervised classification.

We have assumed that the observed trajectories are not corrupted by noise at the discretization points. Although this assumption seems quite reasonable in the case of electricity consumption measurements, it is not true in general. Thus, linear interpolation may not always be effective and linear smoother estimators, such a kernels or smoothing splines, would probably be more appropriate ways to obtain functional versions of the discretized observations.

Finally, another direction for future research is to combine optimal allocation for stratification with model-assisted estimation when auxiliary information is available. There are close relationships between the shape of electricity consumption curves and variables such as past consumption, temperature, household area or type of electricity contract. Such an estimation procedure relies, as noted in Cardot *et al.* (2010), on a parsimonious representation of the trajectories in order to reduce the dimension of the data. One way to achieve this is to first perform a functional principal components analysis and then to model the relationship between the principal components and the auxiliary information.

Acknowledgement

The authors are grateful to the engineers, and particularly to Alain Dessertaine, of the *Département Recherche et Développement* at Électricité de France for fruitful discussions and for allowing us to illustrate this research with the electricity consumption data. This article was also improved by comments and suggestions of the referees as well as discussions with Dr. Camelia Goga and Pauline Lardin. Etienne Josserand thanks the *Conseil Régional de Bourgogne, France* for its financial support (FABER PhD grant).

Appendix : proofs

Proof of Proposition 2.3.1. We study approximation and sampling errors separately :

$$\sup_{t\in[0,T]} |\widehat{\mu}_d(t) - \mu_N(t)| \leq \sup_{t\in[0,T]} |\widehat{\mu}_d(t) - \widehat{\mu}_N(t)| + \sup_{t\in[0,T]} |\widehat{\mu}_N(t) - \mu_N(t)|. \quad (2.12)$$

Suppose $t \in [t_i, t_{i+1}[$ then $|Y_k(t) - \tilde{Y}_k(t)| \leq |Y_k(t_i) - Y_k(t_{i+1})| + |Y_k(t) - Y_k(t_i)|$. By Assumptions 1-2 and an application of the Cauchy–Schwarz inequality,

$$\begin{aligned} |\widehat{\mu}_d(t) - \widehat{\mu}_N(t)| &\leq \frac{1}{N}\sum_{k\in s} \frac{|Y_k(t) - \tilde{Y}_k(t)|}{\pi_k} \\ &\leq \frac{1}{\min_{k\in U_N} \pi_k} \left[\frac{1}{N}\sum_{k\in U} \left\{Y_k(t_i) - \tilde{Y}_k(t)\right\}^2\right]^{1/2} \\ &\leq \frac{1}{\lambda} C_6 |t_{i+1} - t_i|^\beta, \end{aligned}$$

for some positive constant C_6. Consequently,

$$\sqrt{n}\sup_{t\in[0,T]} |\widehat{\mu}_d(t) - \widehat{\mu}_N(t)| \leq \sqrt{n}\frac{C_6}{\lambda} \max_{i\in\{1,\ldots,d_N-1\}} |t_{i+1} - t_i|^\beta. \quad (2.13)$$

We now study the sampling error. Consider the pseudo-metric

$$d_N^2(s,t) = nE\left\{\widehat{\mu}_N(t) - \mu_N(t) - \widehat{\mu}_N(s) + \mu_N(s)\right\}^2$$

for all $(s,t) \in [0,T]\times[0,T]$. We have, for some constant C_7,

$$\begin{aligned} d_N^2(s,t) &\leq \frac{n}{N^2} \sum_{k,\ell\in U_N} \left|\frac{\Delta_{k\ell}}{\pi_k\pi_\ell}\right| |Y_k(t) - Y_k(s)|\,|Y_\ell(t) - Y_\ell(s)| \\ &\leq \frac{n}{N}\frac{C_3}{\lambda}|t-s|^{2\beta} + \frac{n}{\lambda^2}\max_{k\neq\ell}|\Delta_{k\ell}| \left[\frac{1}{N}\sum_{k\in U_N}\left\{Y_k(t) - Y_k(s)\right\}^2\right] \\ &\leq C_7|t-s|^{2\beta}. \end{aligned} \quad (2.14)$$

We apply a result of van der Vaart and Wellner (2000, §2.2) based on maximal inequalities to get the uniform convergence and consider the packing number $D(\epsilon,\ d_N)$, which is the maximum number of points in $[0,T]$ whose distance between each pair is strictly larger

than ϵ. It is clear from (2.14) that $D(\epsilon, d_N) = O(\epsilon^{-1/\beta})$. Considering now the particular Orlicz norm with $\psi(x) = x^2$ in Theorem 2.2.4 of van der Vaart and Welner (2000), we directly final that $\int_0^T \psi^{-1}(\epsilon^{-1/\beta})d\epsilon < \infty$ when $\beta > 1/2$, and consequently there is a constant C_8 such that

$$E\left\{\sqrt{n}\sup_{s,t}|\widehat{\mu}_N(t) - \mu_N(t) - \widehat{\mu}_N(s) + \mu_N(s)|\right\} \leq C_8. \tag{2.15}$$

Since $\sup_t |\widehat{\mu}_N(t) - \mu_N(t)| \leq |\widehat{\mu}_N(0) - \mu_N(0)| + \sup_{s,t} |\widehat{\mu}_N(t) - \mu_N(t) - \widehat{\mu}_N(s) + \mu_N(s)|$, we get the announced result with (2.12), (2.13) and (2.15). □

Proof of Proposition 2.3.2. The proof follows the same lines as the proof of proposition 2.3.1. Let us first write,

$$\sup_{t\in[0,T]}|\widehat{\gamma}_d(t,t) - \gamma_N(t,t)| \leq \sup_{t\in[0,T]}|\widehat{\gamma}_d(t,t) - \widehat{\gamma}_N(t,t)| + \sup_{t\in[0,T]}|\widehat{\gamma}_N(t,t) - \gamma_N(t,t)|$$

Suppose $t \in [t_i, t_{i+1}[$ and define $\delta_{kl}(t) = |\tilde{Y}_l(t) - Y_l(t)|\ |Y_k(t)|$. With Assumptions 1-3, we have, for some constants C_9 and C_{10},

$$\begin{aligned}|\widehat{\gamma}_d(t,t) - \widehat{\gamma}_N(t,t)| &\leq \frac{C_9}{N^2}\left[\sum_{k\in s}\left|\tilde{Y}_k^2(t) - Y_k^2(t)\right| + \max_{k\neq l}|\Delta_{kl}|\sum_{k\in s}\sum_{l\neq k}\{\delta_{kl}(t) + \delta_{lk}(t)\}\right] \\ &\leq \frac{C_{10}}{N}|t_{i+1} - t_i|^\beta.\end{aligned}$$

Thus, using Assumption 1,

$$n\sup_{t\in[0,T]}|\widehat{\gamma}_d(t,t) - \widehat{\gamma}_N(t,t)| \leq C_{10}\max_{i\in\{1,\ldots,d_N-1\}}|t_{i+1} - t_i|^\beta. \tag{2.16}$$

Consider now the sampling error and define, for $(s,t) \in [0,T]\times[0,T]$, $d_N^2(s,t) = n^2 E\left\{\widehat{\gamma}_N(t,t) - \gamma_N(t,t) - \widehat{\gamma}_N(s,s) + \gamma_N(s,s)\right\}^2$ and $\phi_{kl}(s,t) = Y_k(t)Y_l(t) - Y_k(s)Y_l(s)$. We have

$$d_N^2(s,t) = \frac{n^2}{N^4}\sum_{k,l\in U_N}\sum_{k',l'\in U_N}\phi_{kl}(s,t)\phi_{k'l'}(s,t)\frac{\Delta_{kl}}{\pi_k\pi_l}\frac{\Delta_{k'l'}}{\pi_{k'}\pi_{l'}}E\left\{\left(\frac{I_kI_l}{\pi_{kl}} - 1\right)\left(\frac{I_{k'}I_{l'}}{\pi_{k'l'}} - 1\right)\right\}.$$

Following the same lines as the proof of Theorem 3 in Breidt & Opsomer (2000), we get after some algebra that, for some constant C_{11},

$$d_N^2(s,t) \leq C_{11}\left[n^{-1} + \max_{(k,l,k',l')\in D_{4,N}}|E\left\{(I_kI_l - \pi_{kl})(I_{k'}I_{l'} - \pi_{k'l'})\right\}|\right]|t-s|^{2\beta}.$$

Applying again a maximal inequality as in the Proof of Proposition 2.3.1, we get the announced result. □

Proof of Proposition 2.3.3. Noting that, with (2.13), $\sqrt{n}\left\{\widehat{\mu}_d(t) - \mu_N(t)\right\} = \sqrt{n}\left\{\widehat{\mu}_N(t) - \mu_N(t)\right\} + o(1)$, uniformly in t, we only need to study the asymptotic distribution of the random function $X_n(t) = \sqrt{n}\left\{\widehat{\mu}_N(t) - \mu_N(t)\right\}$, for $t \in [0,T]$.

We first consider a m-tuple $(t_1, \ldots, t_m) \in [0,T]^m$, a vector $c^{\mathrm{T}} = (c_1, \ldots, c_m) \in R^m$ and prove that $\sum_{i=1}^m c_i X_n(t_i)$ is asymptotically Gaussian for all $c \in R^m$. Considering $Y_{kc} = \sum_{i=1}^m c_i Y_k(t_i)$, it is clear, with Assumption 6, that $N^{-1} \sum_{k \in U} |Y_{kc}|^{2+\delta} < \infty$ and we have

$$\sum_{i=1}^m c_i X_n(t_i) \quad = \quad \sqrt{n} \left\{ \frac{1}{N} \sum_{k \in s} \frac{Y_{kc}}{\pi_k} - \sum_{i=1}^m c_i \mu_N(t_i) \right\}.$$

Denoting by $\widehat{\mu}_c = N^{-1} \sum_{k \in s} \pi_k^{-1} Y_{kc}$ the Horvitz–Thompson estimator of $\mu_c = N^{-1} \sum_{k \in U} Y_{kc}$, it is clear that $\mu_c = \sum_{i=1}^m c_i \mu_N(t_i)$, $E(\widehat{\mu}_c) = \mu_c$, and with Assumption 6, $\sqrt{n}(\widehat{\mu}_c - E(\widehat{\mu}_c))$ converges in distribution to $N(0, c^{\mathrm{T}} M c)$ where M is a covariance matrix with generic elements $[M]_{ij} = \breve{\gamma}(t_i, t_j)$. The Cramer–Wold device tells us that the vector $(X_n(t_1), \ldots, X_n(t_m))$ is asymptotically multivariate normal.

Secondly, we need to check that X_n satisfies a tightness property in order to get the asymptotic convergence in distribution in the space of continuous functions $C[0,T]$. We have with (2.14), for all $(s,t) \in [0,T] \times [0,T]$, $E\{|X_n(t) - X_n(s)|^2\} \leq C_7 |t-s|^{2\beta}$, and the sequence X_n is tight, when $\beta > 1/2$, according to Theorem 12.3 of Billingsley (1968). □

References

Bickel, P. J. & Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics*, **12**, 470-482.

Billingsley, P. (1968). *Convergence of Probability Measures.* John Wiley, New York.

Breidt, F. J. & Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, **28**, 1026-1053.

Breidt, F. J. & Opsomer, J. D. (2008). Endogeneous post-stratification in surveys: classifying with a sample-fitted model. *Annals of Statistics*, **36**, 403-427.

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

Chen, J. & Rao, J. N. K. (2007). Asymptotic normality under two-phase sampling designs. *Statistica Sinica*, **17**, 1047-1064.

Chiky, R. & Hébrail, G. (2009). Spatio-temporal sampling of distributed data streams. *J. of Computing Science and Engineering*, to appear.

Cochran, W.G. (1977). *Sampling techniques.* 3rd Edition, Wiley, New York.

Degras, D. (2009). Nonparametric estimation of a trend based upon sampled continuous processes. *C. R. Math. Acad. Sci. Paris*, **347**, 191-194.

Degras, D. (2010). Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica*, to appear.

Erdös, P. & Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hungar. Acad. Sci.* **4**, 49-61.

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

Fuller, W.A. (2009). *Sampling Statistics.* John Wiley and Sons.

Hàjek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hungar. Acad. Sci.* **5**, 361-374.

Isaki, C.T. & Fuller, W.A. (1982). Survey design under the regression superpopulation model. *J. Am. Statist. Ass.* **77**, 89-96.

Landau, H. & Shepp, L.A. (1970). On the supremum of a Gaussian process. *Sankhyã*, **32**, 369-378

Müller, H.G. (2005). Functional modelling and classification of longitudinal data (with discussions). *Scand. J. Statist.*, **32**, 223-246.

Ramsay, J. O. & Silverman, B.W. (2005). *Functional Data Analysis.* Springer-Verlag, 2nd ed.

Robinson, P. M. & Särndal, C. E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya : The Indian Journal of Statistics*, **45**, 240-248.

Van der Vaart, A.W. & Wellner, J.A. (2000). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.

Chapitre 3

Compléments

Ce chapitre détaille quelques caractéristiques de l'estimateur de la moyenne dans le cadre du sondage stratifié, ainsi que des démonstrations complètes qui ont soit été omises soit été succinctement développées lors du chapitre 2.

3.1 Estimateur stratifié et allocation optimale

Nous commençons par expliciter la fonction de covariance de notre estimateur de la courbe moyenne $\widehat{\mu}_{\text{strat}}$ dans le cas du sondage stratifié, puis différents critères d'allocation des strates seront abordés.

Proposition 3.1.1. *La fonction de covariance de $\widehat{\mu}_{strat}$, notée $\gamma_{strat}(s,t)$, peut être explicitée par*

$$\gamma_{strat}(s,t) \quad = \quad \frac{1}{N^2}\sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h}\ \widetilde{\gamma}_h(s,t), \quad (s,t) \in [0,T]\times[0,T],$$

avec $\widetilde{\gamma}_h(s,t) = \frac{N_h}{N_h-1}\gamma_h(s,t)$.

Démonstration de la Proposition 3.1.1. On a

$$\begin{aligned}\text{Cov}(\widehat{\mu}_{\text{strat}}(s), \widehat{\mu}_{\text{strat}}(t)) &= \text{Cov}\Big(\frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k\in U_h} Y_k(t)I_k, \frac{1}{N}\sum_{i=1}^{H}\frac{N_i}{n_i}\sum_{l\in U_i} Y_l(s)I_l\Big)\\ &= \frac{1}{N^2}\sum_{h=1}^{H}\sum_{i=1}^{H}\frac{N_h}{n_h}\frac{N_i}{n_i}\sum_{k\in U_h}\sum_{l\in U_i} Y_k(t)Y_l(s)\text{Cov}(I_k, I_l).\end{aligned}$$

Des résultats classiques sur le sondage stratifié nous donne

$$\text{Cov}(I_k, I_l) = \begin{cases} \frac{n_h}{N_h}\frac{N_h-n_h}{N_h} & \text{if } l = k,\ k \in U_h,\\ -\frac{n_h(N_h-n_h)}{N_h^2(N_h-1)} & \text{if } k \text{ and } l \in U_h,\ k \neq l,\\ 0 & \text{if } k \in U_h \text{ and } l \in U_i,\ h \neq i,\end{cases}$$

et ainsi,

$$\begin{aligned}
\mathrm{Cov}(\widehat{\mu}_{\mathrm{strat}}(s),\widehat{\mu}_{\mathrm{strat}}(t)) &= \frac{1}{N^2}\sum_{h=1}^{H}\frac{N_h^2}{n_h^2}\Big(\sum_{k\in U_h}\frac{N_h-n_h}{n_h}Y_k(t)Y_k(s) \\
&\quad -\sum_{k\in U_h}\sum_{l\in U_h l\neq k}\frac{n_h(N_h-n_h)}{N_h^2(N_h-1)}Y_k(t)Y_l(s)\Big) \\
&= \frac{1}{N^2}\sum_{h=1}^{H}\frac{N_h-n_h}{n_h}\Big(\sum_{k\in U_h}Y_k(t)Y_k(s) \\
&\quad -\frac{1}{N_h-1}\sum_{k\in U_h}\sum_{l\in U_h l\neq k}Y_k(t)Y_l(s)\Big) \\
&= \frac{1}{N^2}\sum_{h=1}^{H}N_h\frac{N_h-n_h}{n_h}\Big(\frac{1}{N_h-1}\sum_{k\in U_h}Y_k(t)Y_k(s) \\
&\quad -\frac{N_h}{N_h-1}\mu(t)\mu(s)\Big) \\
&= \frac{1}{N^2}\sum_{h=1}^{H}N_h\frac{N_h-n_h}{n_h}\widetilde{\gamma}_h(s,t).
\end{aligned}$$

□

Une fois la fonction de covariance de notre estimateur obtenue, il est possible de considérer plusieurs critères d'optimisation qui ne sont pas équivalents et qui conduisent à différentes règles d'allocation optimale.

Un premier critère basé sur un point de vue minimax, dans lequel l'objectif est de construire un estimateur dont la variance est aussi petite que possible dans le pire des cas,

$$\min_{(n_1,\ldots,n_H)}\sup_{t\in[0,T]}\mathrm{Var}(\widehat{\mu}_{\mathrm{strat}}(t)) \quad \text{s.t.} \quad \sum_{h=1}^{H}n_h = n \text{ and } n_h>0,\ h=1,\ldots,H. \tag{3.1}$$

Trouver les solutions du problème d'optimisation (3.1) n'est pas trivial et requirerait, entre autre, de connaître les dérivées premières des fonctions de covariance $\gamma_h(t,t)$ dans chaque strate h.

Une seconde approche naturelle serait de minimiser la largeur moyenne des bandes de confiance de l'équation (2.8),

$$\min_{(n_1,\ldots,n_H)}\int_0^T\sqrt{\mathrm{Var}(\widehat{\mu}_{\mathrm{strat}}(t))}\,dt \quad \text{s.t.} \quad \sum_{h=1}^{H}n_h = n \text{ et } n_h>0,\ h=1,\ldots,H. \tag{3.2}$$

Ici aussi, trouver les solutions n'est pas chose aisée. Il n'y a pas de solution explicite, et pour déterminer les allocations optimales il faudrait plutôt utiliser des outils numériques sophistiqués. Néanmoins, la proposition suivante montre que si l'on considère le problème d'optimisation qui consiste à minimiser la variance moyenne au lieu de l'écart-type moyen, on peut alors trouver des solutions simples et explicites.

Proposition 3.1.2. *La solution du problème d'allocation optimale*

$$\min_{(n_1,\ldots,n_H)} \int_0^T Var(\widehat{\mu}_{strat}(t))\ dt \quad s.t. \quad \sum_{h=1}^{H} n_h = n \text{ et } n_h > 0,\ h = 1,\ldots,H.$$

est donnée par

$$n_h^* \ = \ n\ \frac{N_h\sqrt{\int_0^T \widetilde{\gamma}_h(t,t)\ dt}}{\sum_{i=1}^{H} N_i\sqrt{\int_0^T \widetilde{\gamma}_i(t,t)\ dt}}, \quad h = 1,\ldots,H. \tag{3.3}$$

Démonstration de la Proposition 3.1.2. L'idée de la preuve est similaire à celle du cas réel (Cochran, 1977). En introduisant le coefficient de Lagrange λ,

$$\mathscr{L}(n_1,\ldots,n_H,\lambda) = \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} \int_0^T \widetilde{\gamma}_h(t,t)dt + \lambda\left(\sum_{k=1}^{H} n_k - n\right)$$

les solutions du problème d'optimisation doivent satisfaire des conditions sur les dérivées partielles du premier ordre,

$$\frac{\partial \mathscr{L}}{\partial n_h} = -\frac{N_h^2}{n_h^2}\sum_{k=1}^{H}\int_0^T \widetilde{\gamma}_k(t,t)\ dt + \lambda = 0 \text{ pour } h = 1,\ldots,H$$

et

$$\frac{\partial \mathscr{L}}{\partial \lambda} = \sum_{h=1}^{H} n_h - n = 0.$$

Ainsi, on a,

$$n_h = \frac{N_h}{\sqrt{\lambda}}\sqrt{\sum_{k=1}^{K}\int_0^T \widetilde{\gamma}_k(t,t)\ dt} \text{ pour } h = 1,\ldots,H$$

et comme

$$\sum_{h=1}^{H} n_h = n = \frac{\sum_{h=1}^{H} N_h\sqrt{\sum_{k=1}^{K}\int_0^T \widetilde{\gamma}_k(t,t)\ dt}}{\sqrt{\lambda}}$$

on obtient le résultat annoncé,

$$n_h \ = \ n\ \frac{N_h\sqrt{\int_0^T \widetilde{\gamma}_h(t,t)\ dt}}{\sum_{k=1}^{H} N_k\sqrt{\int_0^T \widetilde{\gamma}_k(t,t)\ dt}} \quad \text{pour } h = 1,\ldots,H.$$

□

Les solutions sont analogues au cas multivarié quand on considère un critère de variance totale (Cochran, 1977). Cela signifie qu'une strate qui possède une plus grande variance intra que les autres, aura un taux de sondage n_h/N_h plus important. Avec les poids de sondage optimaux $n_h^*,\ h = 1,\ldots,H$, un calcul direct donne

$$\int_0^T \mathrm{Var}(\widehat{\mu}^*_{\mathrm{strat}}(t))dt \ = \ \frac{1}{N^2}\left(\frac{1}{n}\left(\sum_{h=1}^{H} N_h S_h\right)^2 - \sum_{h=1}^{H} N_h S_h^2\right), \tag{3.4}$$

avec $S_h^2 = \int_0^T \widetilde{\gamma}_h(t,t)dt,\ h = 1, \ldots, H$. Le gain quand on considère l'allocation optimale comparée à l'allocation proportionnelle *i.e.* $n_h = nN_h/N$, est donné par

$$\int_0^T \mathrm{Var}(\widehat{\mu}_{\mathrm{prop}}(t)) - \mathrm{Var}(\widehat{\mu}^*_{\mathrm{strat}}(t))\ dt \quad = \quad \frac{1}{n}\left(\sum_{h=1}^{H} \frac{N_h}{N}\left[S_h - \left(\sum_{\ell=1}^{H} \frac{N_\ell}{N} S_\ell\right)\right]^2\right)$$

où $\widehat{\mu}_{\mathrm{prop}}$ est l'estimateur d'Horvitz-Thompson de la courbe moyenne avec allocation proportionnelle. Ce résultat est à nouveau similaire à celui de Cochran (1977) dans le cas multivarié.

3.2 Consistance des estimateurs de la moyenne et de la covariance

Pour finir, nous présentons en détails les preuves des propositions 2.3.1 et 2.3.2.

Preuve de la Proposition 2.3.1. On étudie séparément l'erreur d'approximation et l'erreur d'échantillonnage :

$$\sup_{t\in[0,T]} |\widehat{\mu}_d(t) - \mu_N(t)| \quad \leq \quad \sup_{t\in[0,T]} |\widehat{\mu}_d(t) - \widehat{\mu}_N(t)| + \sup_{t\in[0,T]} |\widehat{\mu}_N(t) - \mu_N(t)|. \tag{3.5}$$

Pour tout $t \in [0,T]$, l'effet de discrétisation peut être contrôlé en remarquant que si $t \in [t_i, t_{i+1}[$ alors $|Y_k(t) - \tilde{Y}_k(t)| \leq |Y_k(t_i) - Y_k(t_{i+1})| + |Y_k(t) - Y_k(t_i)|$. En utilisant les Hypothèses 1 et 2 ainsi que l'inégalité de Cauchy–Schwarz, on a

$$\begin{aligned}
|\widehat{\mu}_d(t) - \widehat{\mu}_N(t)| &\leq \frac{1}{N}\sum_{k\in s} \frac{|Y_k(t) - \tilde{Y}_k(t)|}{\pi_k} \\
&\leq \frac{1}{\min_{k\in U_N} \pi_k} \frac{1}{N} \sum_{k\in U_N} |Y_k(t) - \tilde{Y}_k(t)| \\
&\leq \frac{1}{\lambda}\left[\frac{1}{N}\sum_{k\in U_N} |Y_k(t_i) - \tilde{Y}_k(t_{i+1})| + \frac{1}{N}\sum_{k\in U_N} |Y_k(t) - \tilde{Y}_k(t_i)|\right] \\
&\leq \frac{1}{\lambda}\left[\left(\frac{1}{N}\sum_{k\in U} \left\{Y_k(t_i) - \tilde{Y}_k(t_{i+1})\right\}^2\right)^{1/2} + \left(\frac{1}{N}\sum_{k\in U} \left\{Y_k(t) - \tilde{Y}_k(t_i)\right\}^2\right)^{1/2}\right] \\
&\leq \frac{1}{\lambda}\left[C_3|t_{i+1} - t_i|^\beta + C_3|t - t_i|^\beta\right] \\
&\leq \frac{1}{\lambda} 2C_3|t_{i+1} - t_i|^\beta \\
&\leq \frac{1}{\lambda} C_6|t_{i+1} - t_i|^\beta,
\end{aligned}$$

pour une certaine constante C_6. Par conséquent,

$$\sqrt{n}\sup_{t\in[0,T]} |\widehat{\mu}_d(t) - \widehat{\mu}_N(t)| \quad \leq \quad \sqrt{n}\frac{C_6}{\lambda} \max_{i\in\{1,\ldots,d_N-1\}} |t_{i+1} - t_i|^\beta. \tag{3.6}$$

Maintenant, on étudie l'erreur d'échantillonnage. Considérons la pseudo-métrique $d_N^2(s,t) = nE\left\{\widehat{\mu}_N(t) - \mu_N(t) - \widehat{\mu}_N(s) + \mu_N(s)\right\}^2$ pour tout $(s,t) \in [0,T]\times[0,T]$. On a, pour une certaine constante C_7,

$$\begin{aligned}
d_N^2(s,t) &\leq \frac{n}{N^2}\sum_{k,\ell\in U_N}\left|\frac{\Delta_{k\ell}}{\pi_k\pi_\ell}\right| |Y_k(t)-Y_k(s)|\,|Y_\ell(t)-Y_\ell(s)| \\
&\leq \frac{n}{N^2}\sum_{k\in U_N}\left|\frac{\Delta_{kk}}{\pi_k^2}\right| |Y_k(t)-Y_k(s)|^2 \\
&\quad + \frac{n}{N^2}\sum_{\substack{k,\ell\in U_N\\ k\neq\ell}}\left|\frac{\Delta_{k\ell}}{\pi_k\pi_\ell}\right| |Y_k(t)-Y_k(s)|\,|Y_\ell(t)-Y_\ell(s)| \\
&\leq \frac{n}{N^2}\sum_{k\in U_N}\left|\frac{\pi_k(1-\pi_k)}{\pi_k^2}\right| |Y_k(t)-Y_k(s)|^2 \\
&\quad + \frac{n}{N^2}\max_{k\neq\ell}|\Delta_{k\ell}|\sum_{\substack{k,\ell\in U_N\\ k\neq\ell}}\left|\frac{1}{\pi_k\pi_\ell}\right| |Y_k(t)-Y_k(s)|\,|Y_\ell(t)-Y_\ell(s)| \\
&\leq \frac{n}{N}\frac{1}{\lambda}\frac{1}{N}\sum_{k\in U_N}|Y_k(t)-Y_k(s)|^2 \\
&\quad + n\max_{k\neq\ell}|\Delta_{k\ell}|\frac{1}{\lambda^2}\left[\frac{1}{N}\sum_{k\in U_N}|Y_k(t)-Y_k(s)|\right]^2 \\
&\leq \frac{n}{N}\frac{C_3}{\lambda}|t-s|^{2\beta} + \frac{n}{\lambda^2}\max_{k\neq\ell}|\Delta_{k\ell}|\left[\frac{1}{N}\sum_{k\in U_N}\left\{Y_k(t)-Y_k(s)\right\}^2\right] \\
&\leq \frac{n}{N}\frac{C_3}{\lambda}|t-s|^{2\beta} + \frac{n}{\lambda^2}\max_{k\neq\ell}|\Delta_{k\ell}|\frac{C_3}{\lambda}|t-s|^{2\beta} \\
&\leq C_7|t-s|^{2\beta}. \qquad (3.7)
\end{aligned}$$

On applique un résultat de van der Vaart and Wellner (2000, §2.2) basé sur les inégalités maximales (voir annexe A.2) afin d'obtenir la convergence uniforme. On considère le *packing number* $D(\epsilon,\ d_N)$ qui est le nombre maximum de points dans $[0,T]$ dont la distance entre chaque paire est strictement plus grande que ϵ. On déduit grâce à (3.7) que $D(\epsilon, d_N) = O(\epsilon^{-1/\beta})$. En considérant à présent le cas particulier de la norme de Orlicz avec $\psi(x) = x^2$ dans le Théorème 2.2.4 de van der Vaart and Welner (2000), on a

$$\int_0^T \psi^{-1}(\epsilon^{-1/\beta})d\epsilon = \int_0^T \epsilon^{-1/2\beta}d\epsilon < \infty \ \text{ quand } \beta > 1/2.$$

Par conséquent, il existe une constante C_8 tel que

$$E\left\{\sqrt{n}\sup_{s,t}|\widehat{\mu}_N(t)-\mu_N(t)-\widehat{\mu}_N(s)+\mu_N(s)|\right\} \ \leq \ C_8. \qquad (3.8)$$

Puisque $\sup_t|\widehat{\mu}_N(t)-\mu_N(t)| \leq |\widehat{\mu}_N(0)-\mu_N(0)| + \sup_{s,t}|\widehat{\mu}_N(t)-\mu_N(t)-\widehat{\mu}_N(s)+\mu_N(s)|$, on obtient le résultat annoncé avec (3.5), (3.6) et (3.8). □

Preuve de la Proposition 2.3.2. La preuve reprend les mêmes étapes que la preuve de la proposition 2.3.1. D'abord, on a

$$\sup_{t\in[0,T]} |\widehat{\gamma}_d(t,t) - \gamma_N(t,t)| \leq \sup_{t\in[0,T]} |\widehat{\gamma}_d(t,t) - \widehat{\gamma}_N(t,t)| + \sup_{t\in[0,T]} |\widehat{\gamma}_N(t,t) - \gamma_N(t,t)| \quad (3.9)$$

Supposons $t \in [t_i, t_{i+1}[$ et définissons $\delta_{kl}(t) = |\tilde{Y}_l(t) - Y_l(t)|\,|Y_k(t)|$. Avec les hypothèses 1-3, on obtient, pour certaines constantes C_9 et C_{10},

$$\begin{aligned}
|\widehat{\gamma}_d(t,t) - \widehat{\gamma}_N(t,t)| &= \left| \frac{1}{N^2} \sum_{k\in s} \sum_{l\in s} \frac{\Delta_{kl}}{\pi_{kl}\pi_k\pi_l} (\tilde{Y}_k(s)\tilde{Y}_l(t) - Y_k(s)Y_l(t)) \right| \\
&\leq \frac{C_9}{N^2} \left[\sum_{k\in s} \left| \tilde{Y}_k^2(t) - Y_k^2(t) \right| + \max_{k\neq l} |\Delta_{kl}| \sum_{k\in s} \sum_{l\neq k} \{\delta_{kl}(t) + \delta_{lk}(t)\} \right] \\
&\leq \frac{C_{10}}{N} |t_{i+1} - t_i|^\beta.
\end{aligned}$$

Ainsi, avec l'hypothèse 1,

$$n \sup_{t\in[0,T]} |\widehat{\gamma}_d(t,t) - \widehat{\gamma}_N(t,t)| \leq C_{10} \max_{i\in\{1,\ldots,d_N-1\}} |t_{i+1} - t_i|^\beta. \quad (3.10)$$

Considérons à présent l'erreur d'échantillonnage et définissons, pour $(s,t) \in [0,T] \times [0,T]$, $d_N^2(s,t) = n^2 E\left\{\widehat{\gamma}_N(t,t) - \gamma_N(t,t) - \widehat{\gamma}_N(s,s) + \gamma_N(s,s)\right\}^2$ et $\phi_{kl}(s,t) = Y_k(t)Y_l(t) - Y_k(s)Y_l(s)$. On a

$$d_N^2(s,t) = \frac{n^2}{N^4} \sum_{k,l\in U_N} \sum_{k',l'\in U_N} \phi_{kl}(s,t)\phi_{k'l'}(s,t) \frac{\Delta_{kl}}{\pi_k\pi_l} \frac{\Delta_{k'l'}}{\pi_{k'}\pi_{l'}} E\left\{ \left(\frac{I_k I_l}{\pi_{kl}} - 1 \right) \left(\frac{I_{k'} I_{l'}}{\pi_{k'l'}} - 1 \right) \right\}.$$

En suivant les mêmes étapes que la preuve du théorème 3 dans Breidt & Opsomer (2000), on obtient après quelques calculs que, pour une certaine constante C_{11},

$$d_N^2(s,t) \leq C_{11} \left[n^{-1} + \max_{(k,l,k',l')\in D_{4,N}} |E\left\{ (I_k I_l - \pi_{kl})(I_{k'} I_{l'} - \pi_{k'l'}) \right\}| \right] |t-s|^{2\beta} \quad (3.11)$$

En effet, on a

$$\begin{aligned}
d_N^2(s,t) &= \frac{n^2}{N^4} \sum_{k=l} \sum_{k'=l'} \phi_{kk}(s,t)\phi_{k'k'}(s,t) \frac{1-\pi_k}{\pi_k^2} \frac{1-\pi_{k'}}{\pi_{k'}^2} E\left\{ (I_k - \pi_k)(I_{k'} - \pi_{k'}) \right\} \\
&\quad + 2\frac{n^2}{N^4} \sum_{k=l} \sum_{k'\neq l'} \phi_{kk}(s,t)\phi_{k'l'}(s,t) \frac{1-\pi_k}{\pi_k^2} \frac{\Delta_{k'l'}}{\pi_{k'}\pi_{l'}\pi_{k'l'}} E\left\{ (I_k - \pi_k)(I_{k'} I_{l'} - \pi_{k'l'}) \right\} \\
&\quad + \frac{n^2}{N^4} \sum_{k\neq l} \sum_{k'\neq l'} \phi_{kl}(s,t)\phi_{k'l'}(s,t) \frac{\Delta_{kl}}{\pi_k\pi_l\pi_{kl}} \frac{\Delta_{k'l'}}{\pi_{k'}\pi_{l'}\pi_{k'l'}} E\left\{ (I_k I_l - \pi_{kl})(I_{k'} I_{l'} - \pi_{k'l'}) \right\} \\
&\leq \frac{n^2}{N^4} \sum_{k\in U_N} \sum_{k'\in U_N} |\phi_{kk}(s,t)\phi_{k'k'}(s,t)| \frac{|\Delta_{kk'}|}{\lambda^4} \\
&\quad + 2\frac{n^2}{N^4} \sum_{k\in U_N} \sum_{k'\neq l'} |\phi_{kk}(s,t)\phi_{k'l'}(s,t)| \frac{|\Delta_{k'l'}|}{\lambda^4\lambda_*} \left| E\left\{ (I_k - \pi_k)(I_{k'} I_{l'} - \pi_{k'l'}) \right\} \right| \\
&\quad + \frac{n^2}{N^4} \sum_{k\neq l} \sum_{k'\neq l'} |\phi_{kl}(s,t)\phi_{k'l'}(s,t)| \frac{|\Delta_{kl}|\,|\Delta_{k'l'}|}{\lambda^4\lambda_*^2} \left| E\left\{ (I_k I_l - \pi_{kl})(I_{k'} I_{l'} - \pi_{k'l'}) \right\} \right| \\
&\leq a_1 + a_2 + a_3,
\end{aligned}$$

$$
\begin{aligned}
a_1 &\leq \frac{n^2}{N^4}\frac{\max_{k\neq k'}|\Delta_{kk'}|}{\lambda^4}\left\{\sum_{k\in U_N}|\phi_{kk}(s,t)|\right\}^2 + \frac{n^2}{N^4}\frac{1}{\lambda^4}\sum_{k\in U_N}|\phi_{kk}(s,t)|^2 \\
&\leq \frac{n^2}{N^2}\frac{\max_{k\neq k'}|\Delta_{kk'}|}{\lambda^4}\left\{\frac{1}{N}\sum_{k\in U_N}\left|Y_k^2(t)-Y_k^2(s)\right|\right\}^2 + \frac{n^2}{N^4}\frac{1}{\lambda^4}\sum_{k\in U_N}\left|Y_k^2(t)-Y_k^2(s)\right|^2 \\
&\leq \frac{n^2}{N^2}\frac{\max_{k\neq k'}|\Delta_{kk'}|}{\lambda^4}\left[\frac{1}{N}\sum_{k\in U_N}\{Y_k(t)-Y_k(s)\}^2\right]\left[\frac{1}{N}\sum_{k\in U_N}\{Y_k(t)+Y_k(s)\}^2\right] \\
&\quad + \frac{n^2}{N^3}\frac{1}{\lambda^4}\left[\frac{1}{N}\sum_{k\in U_N}\{Y_k(t)-Y_k(s)\}^4\right]^{1/2}\left[\frac{1}{N}\sum_{k\in U_N}\{Y_k(t)+Y_k(s)\}^4\right]^{1/2} \\
&\leq \frac{C_{10}}{N}|t-s|^{2\beta}.
\end{aligned}
$$

Après avoir remarqué que

$$
\begin{aligned}
\frac{1}{N^2}\sum_{k',l'\in U_N}|\phi_{k'l'}(s,t)| &\leq \frac{1}{N^2}\sum_{k',l'\in U_N}|Y_{k'}(t)Y_{l'}(t)-Y_{k'}(s)Y_{l'}(s)| \\
&\leq \frac{1}{N^2}\sum_{k',l'\in U_N}|Y_{k'}(t)|\,|Y_{l'}(t)-Y_{l'}(s)| + |Y_{l'}(s)|\,|Y_{k'}(t)-Y_{k'}(s)| \\
&\leq 2\frac{1}{N}\sum_{k'\in U_N}|Y_{k'}(t)|\,\frac{1}{N}\sum_{l'\in U_N}|Y_{l'}(t)-Y_{l'}(s)| \\
&\leq C_{11}|t-s|^{\beta},
\end{aligned}
$$

on a

$$
\begin{aligned}
a_3 &\leq n^2 \max_{(k,l,k',l')\in D_{4,N}}|E\{(I_kI_l-\pi_{kl})(I_{k'}I_{l'}-\pi_{k'l'})\}|\,\frac{\max_{k\neq l}|\Delta_{kl}|^2}{\lambda^4\lambda_*^2}\frac{1}{N^4}\sum_{k\neq l}\sum_{k'\neq l'}|\phi_{kl}(s,t)\phi_{k'l'}(s,t)| \\
&\leq n^2 \max_{(k,l,k',l')\in D_{4,N}}|E\{(I_kI_l-\pi_{kl})(I_{k'}I_{l'}-\pi_{k'l'})\}|\,\frac{\max_{k\neq l}|\Delta_{kl}|^2}{\lambda^4\lambda_*^2}\left\{\frac{1}{N^2}\sum_{k,l\in U_N}|\phi_{kl}(s,t)|\right\}^2 \\
&\leq C_{12}|t-s|^{2\beta}\max_{(k,l,k',l')\in D_{4,N}}|E\{(I_kI_l-\pi_{kl})(I_{k'}I_{l'}-\pi_{k'l'})\}|.
\end{aligned}
$$

En appliquant l'inégalité de Cauchy-Schwarz à a_2 avec a_1 et a_3, on obtient

$$
a_2 \leq C_{13}|t-s|^{2\beta}.
$$

Pour finir, on applique de nouveau une inégalité maximale comme dans la preuve de la proposition 2.3.1 afin d'obtenir le résultat annoncé. □

Chapitre 4

Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data *

Résumé : Lorsque des collections de données fonctionnelles sont trop grandes pour être observées de façon exhaustive, les techniques d'échantillonnage fournissent un moyen efficace pour estimer des quantités globales telles que la fonction moyenne d'une population. En supposant que les données fonctionnelles sont collectées à partir d'une population finie selon un schéma d'échantillonnage probabiliste, avec des mesures discrètes au cours du temps et bruitées, nous proposons d'abord de lisser les trajectoires sélectionnées avec des polynômes locaux et d'ensuite estimer la fonction moyenne grâce à un estimateur d'Horvitz-Thompson. Sous de faibles conditions sur la taille de la population, les temps d'observation, la régularité des trajectoires, le plan d'échantillonnage, la largeur de fenêtre de lissage, nous montrons un Théorème Central Limite dans l'espace des fonctions continues. Nous établissons également la consistance uniforme de l'estimateur de la fonction de covariance et appliquons les résultats précédents pour construire des bandes de confiance globales pour la fonction moyenne. Les bandes atteignent les taux de couverture nominaux et sont obtenues par simulation d'un processus Gaussien conditionnellement à la fonction de covariance estimée. Pour sélectionner la largeur de la fenêtre de lissage, nous proposons une méthode de validation croisée qui tient compte des poids de sondage. Une étude par simulation évalue la performance de notre approche et souligne l'influence du plan d'échantillonnage et du choix de la fenêtre de lissage.

Mots clés : Théorème Central Limite, Données fonctionnelles, Polynômes locaux, Inégalités maximales, Espace des fonctions continues, Supremum de processus Gaussiens, Échantillonnage, Validation croisée pondérée.

∗. Article écrit en collaboration avec Hervé Cardot et David Degras, et soumis pour publication au journal *Bernoulli*

Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data

Hervé Cardot[(a)], David Degras[(b)], Etienne Josserand[(a)]

(a) Institut de Mathématiques de Bourgogne, UMR 5584,
Université de Bourgogne,
9 Avenue Alain Savary, 21078 Dijon, France.
email: {herve.cardot,etienne.josserand}@u-bourgogne.fr

(b) Statistical and Applied Mathematical Sciences Institute,
19 T.W. Alexander Drive, P.O. Box 14006, Research Triangle Park, NC 27709, USA.
email: ddegras@samsi.info

Summary

When collections of functional data are too large to be exhaustively observed, survey sampling techniques provide an effective way to estimate global quantities such as the population mean function. Assuming functional data are collected from a finite population according to a probabilistic sampling scheme, with the measurements being discrete in time and noisy, we propose to first smooth the sampled trajectories with local polynomials and then estimate the mean function with a Horvitz-Thompson estimator. Under mild conditions on the population size, observation times, regularity of the trajectories, sampling scheme, and smoothing bandwidth, we prove a Central Limit Theorem in the space of continuous functions. We also establish the uniform consistency of a covariance function estimator and apply the former results to build global confidence bands for the mean function. The bands attain nominal coverage and are obtained through Gaussian process simulations conditional on the estimated covariance function. To select the bandwidth, we propose a cross-validation method that accounts for the sampling weights. A simulation study assesses the performance of our approach and highlights the influence of the sampling scheme and bandwidth choice.

Keywords: CLT, functional data, local polynomial smoothing, maximal inequalities, space of continuous functions, suprema of Gaussian processes, survey sampling, weighted cross-validation.

4.1 Introduction

The recent development of automated sensors has given access to very large collections of signals sampled at fine time scales. However, exhaustive transmission, storage, and analysis of such massive functional data may incur very large investments. In this context, when the goal is to assess a global indicator like the mean temporal signal, survey sampling techniques are appealing solutions as they offer a good trade-off between statistical accuracy and global cost of the analysis. In particular they are competitive

with signal compression techniques (Chiky and Hébrail, 2008). The previous facts provide some explanation why, although survey sampling and functional data analysis have been long-established statistical fields, motivation for studying them jointly only recently emerged in the literature. In this regard Cardot *et al.* (2010a) examine the theoretical properties of functional principal components analysis (FPCA) in the survey sampling framework. Cardot *et al.* (2010b) harness FPCA for model-assisted estimation by relating the unobserved principal component scores to available auxiliary information. Focusing on sampling schemes, Cardot and Josserand (2011) estimate the mean electricity consumption curve in a population of about 19,000 customers whose electricity meters were read every 30 minutes during one week. Assuming exact measurements, they first perform a linear interpolation of the discretized signals and then consider a functional version of the Horvitz-Thompson estimator. For a fixed sample size, they show that estimation can be greatly improved by utilizing stratified sampling over simple random sampling and they extend the Neyman optimal allocation rule (see *e.g.* Fuller (2009)) to the functional setup. Note however that the finite-sample and asymptotic properties of their estimator rely heavily on the assumption of error-free measurements, which is not always realistic in practice.

The first contribution of the present work is to generalize the framework of Cardot and Josserand (2011) to noisy functional data. Assuming data are observed with errors that may be correlated over time, we replace the interpolation step in their procedure by a data smoothing step based on local polynomials. We extend the previous asymptotic theory by establishing a functional CLT for the resulting mean function estimator and proving the uniform consistency of a related covariance estimator.

In relation to mean function estimation, a key statistical task is to build confidence regions. There exists a vast and still active literature on confidence bands in nonparametric regression. See *e.g.* Sun and Loader (1994), Eubank and Speckman (1993), Claeskens and van Keilegom (2003), Krivobokova *et al.* (2010), and the references therein. When data are functional the literature is much less abundant. One possible approach is to obtain confidence balls for the mean function in a L^2-space. Mas (2007) exploits this idea in a goodness-of-fit test based on the functional sample mean and regularized inverse covariance operator. Using adaptive projection estimators, Bunea *et al.* (2011) build conservative confidence regions for the mean of a Gaussian process. Another approach consists in deriving results in a space C of continuous functions equipped with the supremum norm. This allows to build confidence bands which can be visualized and interpreted as opposed to L^2-confidence balls. It is adopted for example by Faraway (1997) to build bootstrap bands in a varying-coefficients model, by Cuevas *et al.* (2006) to derive various bootstrap bands for functional location parameters, by Degras (2009, 2010) to obtain normal and bootstrap bands using noisy functional data, and by Cardot and Josserand (2011) in the context of a finite population. In the latter work, the strategy was to first establish a CLT in the space C and then derive confidence bands based on a simple but rough approximation to the supremum of a Gaussian process (Landau and Shepp (1970)). Unfortunately, the associated bands depend on the data-generating process only through its variance structure and not its correlation structure, which may cause the empirical coverage to differ from the nominal level. The second innovation of our paper is to propose confidence bands that are easy to implement and attain nominal coverage in the survey sampling/finite population setting. To do so we use Gaussian process simulations as in Cuevas *et al.*

(2006) or Degras (2010). Our contribution is to provide the theoretical underpinning of the construction method, thereby guaranteeing that nominal coverage is attained. The theory we derive involves random entropy numbers, maximal inequalities, and large covariance matrix theory.

Finally, the implementation of the mean function estimator developed in this paper requires to select a bandwidth in the data smoothing step. Objective, data-driven bandwidth selection methods are desirable for this purpose. As explained by Opsomer and Miller (2005), bandwidth selection in the survey estimation context poses specific problems (in particular, the necessity to take the sampling design into account) that make usual cross-validation or mean square error optimization methods inadequate. In view of the model-assisted survey estimation of a population total, these authors propose a cross-validation method that aims at minimizing the variance of the estimator, the bias component being negligible in their setting. In our functional and design-based framework, the bias is however no longer negligible. We therefore devise a novel cross-validation criterion based on weighted least squares, with weights proportional to the sampling weights. For the particular case of simple random sampling without replacement, this criterion reduces to the cross validation technique of Rice and Silverman (1991).

The paper is organized as follows. We fix notations and define our estimators in section 2. In section 3, we introduce our asymptotic framework based on superpopulation models (see Isaki and Fuller, 1982), establish a CLT for the estimator of the mean trajectory in the space of continuous functions, and show the uniform consistency of a covariance function estimator. After that, we prove that by simulating the limiting Gaussian process conditional on its estimated covariance, one can build confidence bands that have asymptotically correct coverage. Simulations are performed in section 4, where different sampling schemes and bandwidth choices are compared to assess the numerical performance of our methodology. The paper ends with a short discussion on topics for future research. Proofs are gathered in an Appendix.

4.2 Notations and estimators

Consider a finite population $U_N = \{1, \dots, N\}$ of size N and suppose that to each unit $k \in U_N$ corresponds a real function X_k on $[0, T]$, with $T < \infty$. We assume that each trajectory X_k belongs to the space of continuous functions $C([0, T])$. Our target is the mean trajectory $\mu_N(t),\ t \in [0, T]$, defined as follows:

$$\mu_N(t) \quad = \quad \frac{1}{N} \sum_{k \in U} X_k(t). \tag{4.1}$$

We consider a random sample s drawn from U_N without replacement according to a fixed-size sampling design $p_N(s)$, where $p_N(s)$ is the probability of drawing the sample s. The size n_N of s is nonrandom and we suppose that the first and second order inclusion probabilities satisfy

- $\pi_k := \mathbb{P}(k \in s) > 0$ for all $k \in U_N$
- $\pi_{kl} := \mathbb{P}(k \& l \in s) > 0$ for all $k, l \in U_N$

so that each unit and each pair of units can be drawn with a non null probability from the population. Note that for simplicity of notation the subscript N has been omitted. Also, by convention, we write $\pi_{kk} = \pi_k$ for all $k \in U_N$.

Assume that noisy measurements of the sampled curves are available at $d = d_N$ fixed discretization points $0 = t_1 < t_2 < \ldots < t_d = T$. For all unit $k \in s$, we observe

$$Y_{jk} = X_k(t_j) + \epsilon_{jk} \tag{4.2}$$

where the measurement errors ϵ_{jk} are centered random variables that are independent across the index k (units) but not necessarily across j (possible temporal dependence). It is also assumed that the random sample s is independent of the noise ϵ_{jk} and the trajectories $X_k(t), t \in [0, T]$ are deterministic.

Our goal is to estimate μ_N as accurately as possible and to build asymptotic confidence bands, as in Degras (2010) and Cardot and Josserand (2011). For this, we must have a uniformly consistent estimator of its covariance function.

4.2.1 Linear smoothers and the Horvitz-Thompson estimator

For each (potentially observed) unit $k \in U_N$, we aim at recovering the curve X_k by smoothing the corresponding discretized trajectory $(Y_{1k}, \ldots, Y_{dk})$ with a linear smoother (e.g. spline, kernel, or local polynomial):

$$\widehat{X}_k(t) = \sum_{j=1}^{d} W_j(t) Y_{jk}. \tag{4.3}$$

Note that the reconstruction can only be performed for the observed units $k \in s$.

Here we use local linear smoothers (see *e.g.* Fan and Gijbels (1997)) because of their wide popularity, good statistical properties, and mathematical convenience. The weight functions $W_j(t)$ can be expressed as

$$W_j(t) = \frac{\frac{1}{dh}\left\{s_2(t) - (t_j - t)s_1(t)\right\} K\left(\frac{t_j - t}{h}\right)}{s_2(t)s_0(t) - s_1^2(t)}, \quad j = 1, \ldots, d, \tag{4.4}$$

where K is a kernel function, $h > 0$ is a bandwidth, and

$$s_l(x) = \frac{1}{dh} \sum_{j=1}^{d} (t_j - t)^l K\left(\frac{t_j - t}{h}\right), \quad l = 0, 1, 2. \tag{4.5}$$

We suppose that the kernel K is nonnegative, has compact support, satisfies $K(0) > 0$ and $|K(s) - K(t)| \leq C|s - t|$ for some finite constant C and for all $s, t \in [0, T]$.

The classical Horvitz-Thompson estimator (see *e.g.* Fuller (2009)) of the mean curve is

$$\begin{aligned} \widehat{\mu}_N(t) &= \frac{1}{N} \sum_{k \in s} \frac{\widehat{X}_k(t)}{\pi_k} \\ &= \frac{1}{N} \sum_{k \in U} \frac{\widehat{X}_k(t)}{\pi_k} I_k, \end{aligned} \tag{4.6}$$

where I_k is the sample membership indicator ($I_k = 1$ if $k \in s$ and $I_k = 0$ otherwise). It holds that $E(I_k) = \pi_k$ and $E(I_k I_l) = \pi_{kl}$.

4.2.2 Covariance estimation

The covariance function of $\widehat{\mu}_N$ can be written as

$$\mathrm{Cov}\left(\widehat{\mu}_N(s), \widehat{\mu}_N(t)\right) = \frac{1}{N}\gamma_N(s,t) \tag{4.7}$$

for all $s, t \in [0,T]$, where

$$\gamma_N(s,t) = \frac{1}{N}\sum_{k,l\in U} \Delta_{kl}\frac{\tilde{X}_k(s)}{\pi_k}\frac{\tilde{X}_l(t)}{\pi_l} + \frac{1}{N}\sum_{k\in U}\frac{1}{\pi_k}E\left(\tilde{\epsilon}_k(s)\tilde{\epsilon}_k(t)\right) \tag{4.8}$$

with

$$\begin{cases} \tilde{X}_k(t) & = \sum_{j=1}^{d} W_j(t) X_k(t_j), \\ \tilde{\epsilon}_k(t) & = \sum_{j=1}^{d} W_j(t)\epsilon_{kj}, \\ \Delta_{kl} & = \mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k\pi_l. \end{cases} \tag{4.9}$$

A natural estimator of $\gamma_N(s,t)$ (see *e.g.* Fuller (2009)) is given by

$$\widehat{\gamma}_N(s,t) = \frac{1}{N}\sum_{k,l\in U}\frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{I_k}{\pi_k}\frac{I_l}{\pi_l}\right)\widehat{X}_k(s)\widehat{X}_l(t). \tag{4.10}$$

It is unbiased and its uniform mean square consistency is established in Section 4.3.2.

4.3 Asymptotic theory

We consider the superpopulation framework introduced by Isaki and Fuller (1982) and discussed in detail by Fuller (2009). Specifically, we study the behaviour of the estimators $\widehat{\mu}_N$ and $\widehat{\gamma}_N$ as population $U_N = \{1, \ldots, N\}$ increases to infinity with N. Recall that the sample size n, inclusion probabilities π_k and π_{kl}, and grid size d all depend on N. In what follows we use the notations c and C for finite, positive constants whose value may vary from place to place. The following assumptions are needed for our asymptotic study.

(A1) (*Sampling design*) $\dfrac{n}{N} \geq c$, $\pi_k \geq c$, $\pi_{kl} \geq c$, and $n|\pi_{kl} - \pi_k\pi_l| \leq C$ for all $k, l \in U_N$ $(k \neq l)$ and $N \geq 1$.

(A2) (*Trajectories*) $|X_k(s) - X_k(t)| \leq C|s-t|^{\beta}$ and $|X_k(0)| \leq C$ for all $k \in U_N$, $N \geq 1$, and $s, t \in [0,T]$, where $\beta > \frac{1}{2}$ is a finite constant.

(A3) (*Growth rates*) $c \leq d(t_{j+1} - t_j) \leq C$ for all $1 \leq j \leq d$, $N \geq 1$, and $\dfrac{d(\log\log N)}{N} \to 0$ as $N \to \infty$.

(A4) (*Measurement errors*) The random vectors $(\epsilon_{k1}, \ldots, \epsilon_{kd})', k \in U_N$, are i.i.d. and follow the multivariate normal distribution with mean zero and covariance matrix $\mathbf{V}_N$. The largest eigenvalue of the covariance matrix satisfies $\|\mathbf{V}_N\| \leq C$ for all $N \geq 1$.

Assumption (A1) deals with the properties of the sampling design. It states that the sample size must be at least a positive fraction of the population size, that the one- and two-fold inclusion probabilities must be larger than a positive number, and that the two-fold inclusion probabilities should not be to far from independence. The latter is fulfilled

for example for stratified sampling with sampling without replacement within each stratum (Robinson and Särndal (1983)). Assumption (A2) imposes Hölder continuity on the trajectories, a mild regularity condition. Assumption (A3) states that the design points have a quasi-uniform repartition (this holds in particular for equidistant designs and designs generated by a regular density function) and that the grid size is essentially negligible compared to the population size (for example if $d_N \propto N^\alpha$ for some $\alpha \in (0,1)$). In fact the results of this paper also hold if d_N/N stays bounded away from zero and infinity as $N \to \infty$ (see Section 5). Finally (A4) imposes joint normality, short range temporal dependence, and bounded variance for the measurement errors $\epsilon_{kj}, 1 \leq j \leq d$. It is trivially satisfied if the $\epsilon_{kj} \sim N(0, \sigma_j^2)$ are independent with variances $\mathrm{Var}(\epsilon_{kj}) \leq C$. It is also verified if the ϵ_{kj} arise from a discrete time Gaussian process with short term temporal correlation such as ARMA or stationary mixing processes. Note that the Gaussian assumption is not central to our derivations: it can be weakened and replaced by moment conditions on the error distributions at the expense of much more complicated proofs.

4.3.1 Limit distribution of the Horvitz-Thompson estimator

Proceeding further, we would now like to derive the asymptotic distribution of our estimator $\widehat{\mu}_N$ in order to build asymptotic confidence intervals and bands. Obtaining the asymptotic normality of estimators in survey sampling is a technical and difficult issue even for simple quantities such as means or totals of real numbers. Although confidence intervals are commonly used in the survey sampling community, the Central Limit Theorem (CLT) has only been checked rigorously, as far as we know, for a few sampling designs. Erdös and Rényi (1959) and Hàjek (1960) proved that the Horvitz-Thompson estimator is asymptotically Gaussian for simple random sampling without replacement. These results were extended more recently to stratified sampling (Bickel and Freedman (1994)) and some particular cases of two-phase sampling designs (Chen and Rao (2007)). Let us assume that the Horvitz-Thompson estimator satisfies a CLT for real valued quantities.

(A5) (*Univariate CLT*) For any fixed $t \in [0,T]$, it holds that

$$\frac{\widehat{\mu}_N(t) - \mu_N(t)}{\sqrt{\mathrm{Var}(\widehat{\mu}_N(t))}} \rightsquigarrow N(0,1)$$

as $N \to \infty$, where $\rightsquigarrow$ stands for convergence in distribution.

We recall here the definition of the weak convergence in $C([0,T])$ equipped with the supremum norm $\|\cdot\|_\infty$ (e.g. van der Vaart and Wellner (2000)). A sequence (ξ_N) of random elements of $C([0,T])$ is said to converge weakly to a limit ξ in $C([0,T])$ if $\mathbb{E}(\phi(\xi_N)) \to \mathbb{E}(\phi(\xi))$ as $N \to \infty$ for all bounded, uniformly continuous functional ϕ on $(C([0,T]), \|\cdot\|_\infty)$.

To establish the limit distribution of $\widehat{\mu}_N$ in $C([0,T])$, we need to assume the existence of a limit covariance function

$$\gamma(s,t) = \lim_{N\to\infty} \frac{1}{N} \sum_{k,l \in U_N} \Delta_{kl} \frac{X_k(s)}{\pi_k} \frac{X_l(t)}{\pi_l}\,.$$

In the following theorem we state the asymptotic normality of the estimator $\widehat{\mu}_N$ in the space $C([0,T])$ equipped with the sup norm.

Theorem 4.3.1. *Assume (A1)–(A5) and that* $\sqrt{N}h^{\beta} \to 0$ *and* $dh/\log d \to \infty$ *as* $N \to \infty$. *Then*

$$\sqrt{N}\,(\widehat{\mu}_N - \mu_N) \rightsquigarrow G$$

in $C([0,T])$, *where* G *is a Gaussian process with mean zero and covariance function* γ.

Theorem 4.3.1 provides a convenient way to infer the local features of μ_N. It is applied in Section 4.3.3 to the construction of simultaneous confidence bands, but it can also be used for a variety of statistical tests based on supremum norms (see Degras (2010)).

Observe that the conditions on the bandwidth h and design size d are not very constraining. Suppose for example that $d \propto N^{\eta}$ and $h \propto N^{-\nu}$ for some $\eta, \nu > 0$. Then d and h satisfy the conditions of Theorem 4.3.1 as soon as $(2\beta)^{-1} < \nu < \eta < 1$. Thus, for more regular trajectories, *i.e.* larger β, the bandwidth h can be chosen with more flexibility.

The proof of Theorem 4.3.1 is similar in spirit to that of Theorem 1 in Degras (2010) and Proposition 3 in Cardot and Josserand (2011). Essentially, it breaks down into: (i) controlling uniformly on $[0,T]$ the bias of $\widehat{\mu}_N$, (ii) establishing the functional asymptotic normality of the local linear smoother applied to the sampled curves X_k, and (iii) controlling uniformly on $[0,T]$ (in probability) the local linear smoother applied to the errors ϵ_{jk}. Part (i) is easily handled with standard results on approximation properties of local polynomial estimators (see *e.g.* Tsybakov (2009)). Part (ii) mainly consists in proving an asymptotic tightness property, which entails the computation of entropy numbers and the use of maximal inequalities (see van der Vaart and Wellner (2000)). Part (iii) requires first to show the finite-dimensional convergence of the smoothed error process to zero and then to establish its tightness with similar arguments as in part (ii).

4.3.2 Uniform consistency of the covariance estimator

We first note that under (A1)–(A4), by the approximation properties of local linear smoothers, γ_N converges uniformly to γ on $[0,T]^2$ as $h \to 0$ and $N \to \infty$. Hence the consistency of $\widehat{\gamma}_N$ can be stated with respect to γ instead of γ_N. In alignment with the related Proposition 2 in Cardot and Josserand (2010) and Theorem 3 in Breidt and Opsomer (2000), we need to make some assumption on the two-fold inclusion probabilities of the sampling design p_N:

(A6)

$$\lim_{N\to\infty} \max_{(k_1,k_2,k_3,k_4)\in D_{4,N}} \left|\mathbb{E}\{(I_{k_1}I_{k_2} - \pi_{k_1k_2})(I_{k_3}I_{k_4} - \pi_{k_3k_4})\}\right| = 0$$

where $D_{4,N}$ is the set of all quadruples (k_1,k_2,k_3,k_4) in U_N with distinct elements.

This assumption is discussed in detail in Breidt and Opsomer (2000) and is fulfilled for example for stratified sampling.

Theorem 4.3.2. *Assume (A1)–(A4), (A6), and that* $h \to 0$ *and* $dh^{1+\alpha} \to \infty$ *for some* $\alpha > 0$ *as* $N \to \infty$. *Then*

$$\lim_{N\to\infty} E\left(\sup_{s,t\in[0,T]^2} \left|\widehat{\gamma}_N(s,t) - \gamma(s,t)\right|^2\right) = 0.$$

Note the additional condition on the bandwidth h in Theorem 2. If we suppose, as in the remark in Section 4.3.1, that $d \propto N^{\eta}$ and $h \propto N^{-\nu}$ for some $(2\beta)^{-1} < \nu < \eta < 1$, then condition $dh^{1+\alpha} \to \infty$ as $N \to \infty$ is fulfilled with *e.g.* $\alpha = 1 - \eta/2\nu$.

4.3.3 Global confidence bands

In this section we build global confidence bands for μ_N of the form

$$\left\{ \left[\widehat{\mu}_N(t) \pm c \, \frac{\widehat{\sigma}_N(t)}{N^{1/2}} \right], \, t \in [0,T] \right\}, \tag{4.11}$$

where c is a suitable number and $\widehat{\sigma}_N(t) = \widehat{\gamma}_N(t,t)^{1/2}$. More precisely, given a confidence level $1-\alpha \in (0,1)$, we seek $c = c_\alpha$ that approximately satisfies

$$\mathbb{P}\left(|G(t)| \leq c \, \sigma(t), \, \forall t \in [0,T] \right) = 1-\alpha, \tag{4.12}$$

where G is a Gaussian process with mean zero and covariance function γ, and where $\sigma(t) = \gamma(t,t)^{1/2}$. Exact bounds for the supremum of Gaussian processes have only been derived for only a few particular cases (Adler and Taylor, 2007, Chapter 4). Computing accurate and as explicit as possible bounds in a general setting is a difficult issue and would require additional strong conditions such as stationarity which have no reason to be fulfilled in our setting.

In view of Theorems 4.3.1-4.3.2 and Slutski's Theorem, the bands defined in (4.11) with c chosen as in (4.12) will have approximate coverage level $1-\alpha$. The following result provides a simulation-based method to compute c.

Theorem 4.3.3. *Assume (A1)–(A6) and $dh^{1+\alpha} \to \infty$ for some $\alpha > 0$ as $N \to \infty$. Let G be a Gaussian process with mean zero and covariance function γ. Let $(\widehat{G}_N)$ be a sequence of processes such that for each N, conditionally on $\widehat{\gamma}_N$, $\widehat{G}_N$ is Gaussian with mean zero and covariance $\widehat{\gamma}_N$ defined in* (4.10)*. Then for all $c > 0$, as $N \to \infty$, the following convergence holds in probability:*

$$\mathbb{P}\left(|\widehat{G}_N(t)| \leq c \, \widehat{\sigma}_N(t), \, \forall t \in [0,T] \,|\, \widehat{\gamma}_N \right) \to \mathbb{P}\left(|G(t)| \leq c \, \sigma(t), \, \forall t \in [0,T] \right).$$

Theorem 4.3.3 is derived by showing the weak convergence of $(\widehat{G}_N)$ to G in $C([0,T])$, which stems from Theorem 4.3.2 and the Gaussian nature of the processes $\widehat{G}_N$. As in the first two theorems, maximal inequalities are used to obtain the above weak convergence. The practical importance of Theorem 4.3.3 is that it allows to estimate the number c in (4.12) via simulation: (with the previous notations), conditionally on $\widehat{\gamma}_N$, one can simulate a large number of sample paths of the Gaussian process $(\widehat{G}_N/\widehat{\sigma}_N)$ and compute their supremum norms. One then obtains a precise approximation to the distribution of $\|\widehat{G}_N/\widehat{\sigma}_N\|_\infty$, and it suffices to set c as the quantile of order $(1-\alpha)$ of this distribution:

$$\mathbb{P}\left(|\widehat{G}_N(t)| \leq c \, \widehat{\sigma}_N(t), \, \forall t \in [0,T] \,|\, \widehat{\gamma}_N \right) = 1-\alpha. \tag{4.13}$$

Corollary 4.3.1. *Assume (A1)–(A6). Under the conditions of Theorems 4.3.1-4.3.2-4.3.3, the bands defined in* (4.11) *with the real $c = c(\widehat{\gamma}_N)$ chosen as in* (4.13) *have asymptotic coverage level $1-\alpha$, i.e.*

$$\lim_{N\to\infty} \mathbb{P}\left(\mu_N(t) \in \left[\widehat{\mu}_N(t) \pm c \, \frac{\widehat{\sigma}_N(t)}{N^{1/2}} \right], \quad \forall t \in [0,T] \right) = 1-\alpha.$$

Stratum number	1	2	3	4	5
Stratum size	10000	4000	3000	2000	1000
Allocation	655	132	98	68	47

Table 4.1: Strata sizes and optimal allocations.

4.4 A simulation study

In this section, we evaluate the performances of the mean curve estimator as well as the coverage and the width of the confidence bands for different bandwidth selection criteria and different levels of noise. The simulations are conducted in the R environment.

4.4.1 Simulated data and sampling designs

We have generated a population of $N = 20000$ curves discretized at $d = 200$ and $d = 400$ equidistant instants of time in $[0, 1]$. The curves of the population are generated so that they have approximately the same distribution as the electricity consumption curves analyzed in Cardot & Josserand (2011) and each individual curve X_k, for $k \in U$, is simulated as follows

$$X_k(t) = \mu(t) + \sum_{\ell=1}^{3} Z_\ell \; v_\ell(t), \quad t \in [0, 1], \tag{4.14}$$

where the mean function μ is drawn in Figure 4.2 and the random variables Z_ℓ are independent realizations of a centered Gaussian random variable with variance σ_ℓ^2. The three basis function v_1, v_2 and v_3 are orthonormal functions which represent the main mode of variation of the signals, they are represented in Figure 4.1. Thus, the covariance function of the population $\gamma(s, t)$ is simply

$$\gamma(s,t) \quad = \quad \sum_{\ell=1}^{3} \sigma_\ell^2 \; v_\ell(s) v_\ell(t). \tag{4.15}$$

To select the samples, we have considered two probabilistic selection procedures, with fixed sample size, $n = 1000$,

- Simple random sampling without replacement (SRSWR).
- Stratified sampling with SRSWR in all strata. The population U is divided into a fixed number of $G = 5$ strata built by considering the quantiles $q_{0.5}, q_{0.7}, q_{0.85}$ and $q_{0.95}$ of the total consumption $\int_0^1 X_k(t)dt$ for all units $k \in U$. For example, the first strata contains all the units k such that $\int_0^1 X_k(t)dt \leq q_{0.5}$, and thus its size is half of the population size N. The sample size n_g in stratum g is determined by a Neyman-like allocation, as suggested in Cardot and Josserand (2011), in order to get a Horvitz-Thompson estimator of the mean trajectory whose variance is as small as possible. The sizes of the different strata, which are optimal according to this mean variance criterion, are reported in Table 4.1.

We suppose we observe, for each unit k in the sample s, the discretized trajectories, at d equispaced points, $0 = t_1 < \ldots < t_d = 1$,

$$Y_{jk} = X_k(t_j) + \delta \epsilon_{jk} \tag{4.16}$$

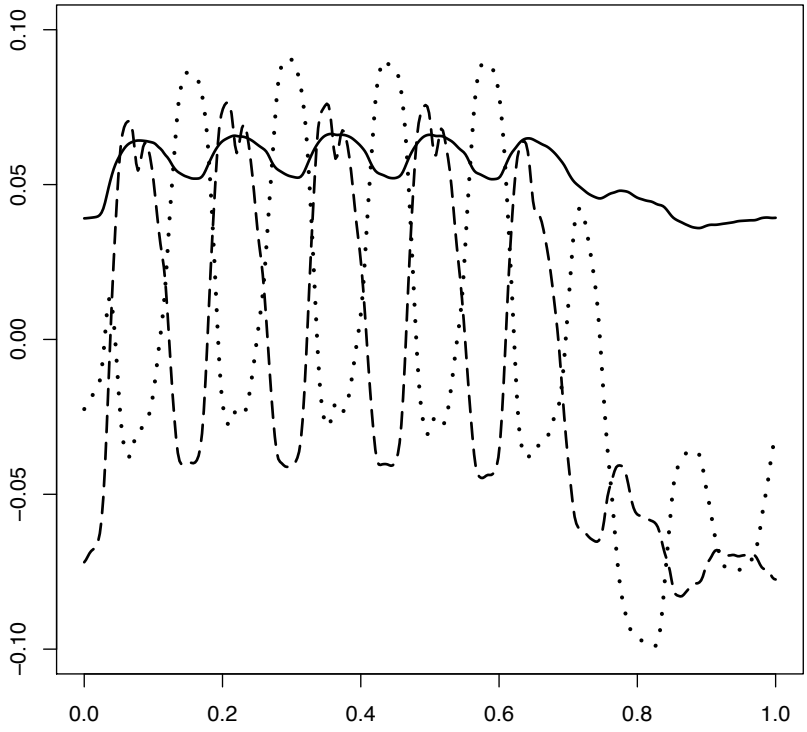

Figure 4.1: Basis functions v_1 (solid line), v_2 (dashed line) and v_3 (dotted line).

where the $\epsilon_{jk} \sim N(0, \gamma(t_j, t_j))$ are independent random variables and the parameter δ allows to control the noise level. As an illustrative example, a sample of $n = 10$ noisy discretized curves are plotted in Figure 4.2.

4.4.2 Weighted cross-validation for bandwidth selection

Assuming we can access the exact trajectories X_k, $k \in s$, (which is the case in simulations) we consider the oracle-type estimator

$$\widehat{\mu}_s = \sum_{k \in s} \frac{X_k}{\pi_k}\,, \tag{4.17}$$

which will be a benchmark in our numerical study. We compare different interpolation and smoothing strategies for estimating the X_k, $k \in s$:

– Linear interpolation of the Y_{jk} as in Cardot and Josserand (2011).

– Local linear smoothing of the Y_{jk} with bandwidth h as in (4.3).

The crucial element here is h. To evaluate the interest of smoothing and the performances of data-driven bandwidth selection criteria, we consider an error measure that compares

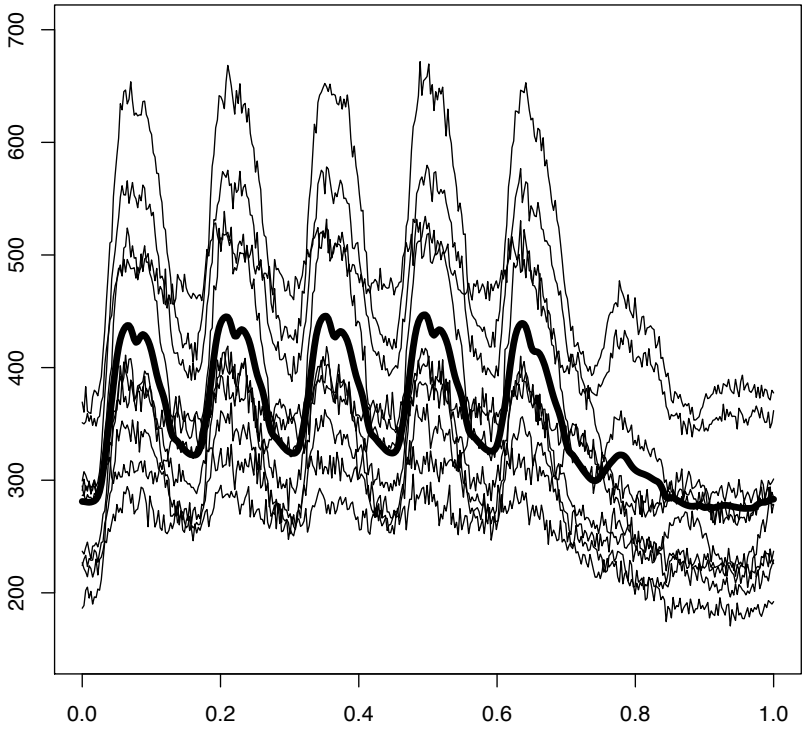

Figure 4.2: A sample of 10 curves for $\delta = 0.05$. The mean profile is plotted in bold line.

the oracle $\widehat{\mu}_s$ to any estimator $\widehat{\mu}$ based on the noisy data Y_{jk}, $k \in s$, $j = 1, \dots d$:

$$L(\widehat{\mu}) = \int_0^T (\widehat{\mu}_s(t) - \widehat{\mu}(t))^2 \, dt \,. \tag{4.18}$$

Considering the estimator defined in (4.6), we denote by h_{oracle} the bandwidth h that minimizes (4.18) and call smooth oracle the corresponding estimator.

When $\sum_{k \in s} \pi_k^{-1} = N$, as in SRSWR and stratified sampling, it can be easily checked that $\widehat{\mu}_s$ is the minimum argument of the weighted least squares functional

$$\sum_{k \in s} w_k \int_0^T (X_k(t) - \mu(t))^2 \, dt \tag{4.19}$$

with respect to $\mu \in L^2([0,T])$, where the weights are $w_k = (N\pi_k)^{-1}$. Then, a simple and natural way to select bandwidth h is to consider the following design-based cross validation

$$\text{WCV}(h) = \sum_{k \in s} w_k \sum_{j=1}^{d} \left(Y_{jk} - \widehat{\mu}_N^{-k}(t_j) \right)^2 . \tag{4.20}$$

where

$$\widehat{\mu}_N^{-k}(t) = \sum_{\ell \in s, \ell \neq k} \widetilde{w}_\ell \ \widehat{X}_\ell(t),$$

with new weights $\widetilde{w}_\ell$. A heuristic justification for this approach is that, given s, we have $E\left[\epsilon_{jk}(X_k(t_j) - \widehat{\mu}_N^{-k}(t_j))|s\right] = 0$ for $j = 1, \ldots, d$ and $k \in s$. Thus,

$$\begin{aligned} E\left[\mathrm{WCV}(h)|s\right] &= \sum_{k \in s} w_k \sum_{j=1}^{d} \left\{ E\left[\left(X_k(t_j) - \widehat{\mu}_N^{-k}(t_j)\right)^2 |s\right] + 2E\left[\epsilon_{jk}(X_k(t_j) - \widehat{\mu}_N^{-k}(t_j))|s\right] + E\left[\epsilon_{jk}^2\right] \right\} \\ &= \sum_{k \in s} w_k \sum_{j=1}^{d} E\left[\left(X_k(t_j) - \widehat{\mu}_N^{-k}(t_j)\right)^2 |s\right] + \mathrm{tr}(\mathbf{V}_N) \end{aligned}$$

and, up to $\mathrm{tr}(\mathbf{V}_N)$ which does not depend on h, the minimum value of the expected cross validation criterion should be attained for estimators which are not too far from $\widehat{\mu}_s$.

This weighted cross validation criterion is simpler than the cross validation criteria based on the estimated variance proposed in Opsomer and Miller (2005). Indeed, in our case, the bias may be non negligible and focusing only on the variance part of the error leads to too large selected values for the bandwidth. Furthermore, Opsomer and Miller (2005) suggested to consider weights defined as follows $\widetilde{w}_\ell = w_\ell/(1 - w_k)$. For SRSWR, since $w_k = n^{-1}$ one has $\widetilde{w}_k = (n-1)^{-1}$, so that the weighted cross validation criterion defined in (4.20) is exactly the cross validation criterion introduced by Rice and Silverman (1991) in the independent case. We denote in the following by h_{cv} the bandwidth value minimizing this criterion.

For stratified sampling, a better approximation which keeps the design-based properties of the estimator $\widehat{\mu}_N^{-k}$ can be obtained by taking into account the sampling rates in the different strata. We have G strata with sizes N_g, $g = 1, \ldots, G$ and we sample n_g observations, with SRSWR, in each stratum g. If unit k comes from strata g, we have $w_k = N_g(Nn_g)^{-1}$. Thus, we take $\widetilde{w}_\ell = (N_g - 1)\{(N-1)(n_g-1)\}^{-1}$ for all the units $\ell \neq k$ in stratum g and just scale the weights for all the units ℓ' of the sample that do not belong to stratum g, $\widetilde{w}_{\ell'} = N(N-1)^{-1}w_{\ell'}$. We denote by h_{wcv} the bandwidth value minimizing (4.20).

4.4.3 Estimation errors and confidence bands

We draw 1000 samples in the population of curves and compare the different estimators of Section 4.2 with the L^2 loss criterion

$$R(\widehat{\mu}) = \int_0^T \left(\widehat{\mu}(t) - \mu(t)\right)^2 dt \tag{4.21}$$

for different values of δ and d in (4.16).

The empirical mean as well as the first, second and third quartiles of the estimation error $R(\widehat{\mu})$ are given in Table 4.2 for $d = 200$ and in Table 4.3 for $d = 400$. We can first note that stratified sampling allows to improve much the estimation of the mean curve. We also remark that, for such large samples, linear interpolation performs nearly as well as the smooth oracle estimator, especially when the noise level is low ($\delta \leq 15\%$). As far

		SRSWR				Stratified sampling			
δ	h	Mean	1Q	Median	3Q	Mean	1Q	Median	3Q
5%	lin	17.65	3.078	8.729	23.50	4.221	1.444	2.791	5.594
	h_{cv}	17.65	3.066	8.710	23.51	6.493	3.605	5.356	8.028
	h_{wcv}	17.65	3.066	8.710	23.51	4.221	1.448	2.781	5.555
	h_{oracle}	17.65	3.068	8.725	23.50	4.220	1.446	2.778	5.571
	$\widehat{\mu}_s$	17.60	3.011	8.698	23.36	4.174	1.378	2.758	5.548
10%	lin	17.18	3.226	9.019	22.20	4.335	1.535	3.040	5.675
	h_{cv}	17.17	3.201	8.975	22.26	6.688	3.699	5.354	8.091
	h_{wcv}	17.17	3.201	8.975	22.26	4.342	1.613	3.068	5.637
	h_{oracle}	17.17	3.209	8.969	22.23	4.330	1.573	3.053	5.627
	$\widehat{\mu}_s$	17.00	3.092	8.780	22.11	4.136	1.359	2.835	5.486
15%	lin	18.08	3.616	9.589	23.58	4.263	1.755	3.050	5.614
	h_{cv}	18.06	3.633	9.557	23.44	6.473	3.641	5.336	8.064
	h_{wcv}	18.06	3.633	9.557	23.44	4.238	1.703	3.041	5.682
	h_{oracle}	18.06	3.634	9.568	23.43	4.225	1.702	3.012	5.631
	$\widehat{\mu}_s$	17.69	3.282	9.263	22.77	3.812	1.307	2.625	5.131
20%	lin	16.98	3.722	9.222	21.31	4.870	2.187	3.755	6.047
	h_{cv}	16.91	3.657	9.226	21.28	7.025	4.022	5.878	8.838
	h_{wcv}	16.91	3.657	9.226	21.28	4.791	2.110	3.683	6.014
	h_{oracle}	16.90	3.668	9.221	21.29	4.779	2.110	3.681	6.000
	$\widehat{\mu}_s$	16.27	3.040	8.606	20.71	4.086	1.373	2.964	5.283
25%	lin	17.69	3.940	8.989	21.52	5.257	2.625	4.148	6.535
	h_{cv}	17.53	3.826	8.755	21.53	6.982	4.287	5.829	8.469
	h_{wcv}	17.53	3.826	8.755	21.53	5.017	2.388	3.893	6.331
	h_{oracle}	17.52	3.806	8.778	21.52	5.007	2.369	3.883	6.269
	$\widehat{\mu}_s$	16.58	2.847	7.870	20.01	4.069	1.457	2.944	5.278

Table 4.2: Estimation errors according to $R(\widehat{\mu})$ for different noise levels and bandwidth values, with $d = 200$ time instants. Units are selected by simple random sampling without replacements (SRSWR) or stratified sampling.

as bandwidth selection is concerned, we can note that the usual cross validation criterion h_{cv} is not adapted to unequal probability sampling and does not perform as well as linear interpolation for stratified sampling by selecting too large values for the bandwidth. On the other hand, the weighted cross-validation criterion seems to be effective to select good bandwidth values and produce estimators whose estimation errors are very close to the oracle and perform better than the other estimators when the noise level is moderate or high ($\delta \geq 20\%$).

This is clearer when we look at criterion $L(\widehat{\mu})$, defined in (4.18), which only focus on the part of the estimation error which is due to the noise. Results are presented in Table 4.4 for $d = 200$ and in Table 4.5 for $d = 400$. We can also note that there is a significant effect of the number of discretization points on the accuracy of the smoothed estimators. Our individual trajectories, which have roughly the same shape as load curves, are actually not very smooth so that smoothing approaches are only really interesting, compared to linear interpolation when the number of discretization points d is large enough.

We now examine in Table 4.6 and Table 4.7 the empirical coverage and the width of the confidence bands, which are built as described in Section 4.3.3. For each sample, we estimate the covariance function $\widehat{\gamma}_N$ and draw 10000 realizations of a centered Gaus-

		SRSWR				Stratified sampling			
δ	h	Mean	1Q	Median	3Q	Mean	1Q	Median	3Q
5%	lin	18.03	3.388	9.243	23.27	4.049	1.446	2.858	5.353
	h_{cv}	18.02	3.384	9.257	23.34	6.092	3.244	4.868	7.555
	h_{wcv}	18.02	3.384	9.257	23.34	4.047	1.449	2.821	5.398
	h_{oracle}	18.02	3.387	9.269	23.32	4.043	1.433	2.828	5.388
	$\widehat{\mu}_s$	17.98	3.353	9.199	23.17	4.000	1.388	2.809	5.294
10%	lin	16.97	3.084	8.058	21.46	4.294	1.683	3.208	5.797
	h_{cv}	16.93	2.979	7.916	21.45	6.207	3.308	5.137	7.800
	h_{wcv}	16.93	2.979	7.916	21.45	4.233	1.577	3.123	5.741
	h_{oracle}	16.93	2.970	7.914	21.44	4.229	1.579	3.118	5.703
	$\widehat{\mu}_s$	16.81	2.876	7.811	21.45	4.099	1.512	3.006	5.608
15%	lin	19.03	3.761	10.09	24.80	4.528	1.772	3.367	5.994
	h_{cv}	18.87	3.642	9.899	24.54	6.259	3.446	5.327	7.829
	h_{wcv}	18.87	3.642	9.899	24.54	4.335	1.630	3.188	5.828
	h_{oracle}	18.87	3.642	9.888	24.52	4.330	1.612	3.178	5.826
	$\widehat{\mu}_s$	18.61	3.414	9.665	24.24	4.080	1.340	2.918	5.538
20%	lin	17.06	3.635	8.545	22.95	4.749	2.128	3.643	6.144
	h_{cv}	16.69	3.288	8.060	22.82	6.362	3.489	5.205	8.009
	h_{wcv}	16.69	3.288	8.060	22.82	4.353	1.755	3.231	5.675
	h_{oracle}	16.69	3.267	8.044	22.77	4.347	1.734	3.216	5.694
	$\widehat{\mu}_s$	16.35	2.993	7.860	22.13	3.960	1.333	2.867	5.414
25%	lin	18.16	3.885	9.427	22.86	5.254	2.845	4.236	6.572
	h_{cv}	17.55	3.303	8.889	22.09	6.452	3.770	5.372	8.114
	h_{wcv}	17.55	3.303	8.889	22.09	4.566	2.121	3.486	5.809
	h_{oracle}	17.55	3.282	8.898	22.09	4.561	2.107	3.477	5.812
	$\widehat{\mu}_s$	17.04	2.750	8.383	21.87	4.043	1.602	3.018	5.306

Table 4.3: Estimation errors according to $R(\widehat{\mu})$ for different noise levels and bandwidth values, with $d = 400$ time instants. Units are selected by simple random sampling without replacements (SRSWR) or stratified sampling.

sian process with variance function $\widehat{\gamma}_N$ in order to obtain a suitable coefficient c with a confidence level of $1 - \alpha = 0.95$ as explained in equation (4.13). The area of the confidence band is then $\int_0^T 2c\sqrt{\widehat{\gamma}(t,t)}\ \mathrm{d}t$. The results highlight now the interest of considering smoothing strategies combined with the weighted cross validation bandwidth selection criterion (4.20). It appears that linear interpolation, which does not intend to get rid of the noise, always gives larger confidence bands than the smoothed estimators based on h_{wcv}. As before, smoothing approaches become more interesting as the number of discretization points and the noise level increase. The empirical coverage of the smoothed estimator is lower than the linear interpolation estimator but remains slightly higher than the nominal one.

As a conclusion of this simulation study, it appears that smoothing is not a crucial aspect when the only target is the estimation of the mean, and that bandwidth values should be chosen by a cross validation criterion that takes the sampling weights into account. When the goal is also to build confidence bands, smoothing with weighted cross validation criteria lead to narrower bands compared to interpolation techniques, without deteriorating the empirical coverage.

		SRSWR				Stratified sampling			
δ	h	Mean	1Q	Median	3Q	Mean	1Q	Median	3Q
5%	lin	0.044	0.041	0.044	0.047	0.049	0.046	0.049	0.053
	h_{cv}	0.044	0.041	0.044	0.048	2.520	2.083	2.852	3.032
	h_{wcv}	0.044	0.041	0.044	0.048	0.058	0.054	0.058	0.062
	h_{oracle}	0.044	0.041	0.044	0.047	0.049	0.045	0.049	0.052
10%	lin	0.175	0.163	0.175	0.186	0.196	0.181	0.194	0.208
	h_{cv}	0.170	0.158	0.170	0.182	2.626	2.162	2.902	3.110
	h_{wcv}	0.170	0.158	0.170	0.182	0.202	0.188	0.201	0.216
	h_{oracle}	0.169	0.156	0.168	0.180	0.188	0.174	0.187	0.200
15%	lin	0.396	0.366	0.394	0.424	0.443	0.411	0.440	0.474
	h_{cv}	0.368	0.342	0.366	0.394	2.743	2.264	2.972	3.206
	h_{wcv}	0.368	0.342	0.366	0.394	0.417	0.388	0.413	0.446
	h_{oracle}	0.365	0.339	0.364	0.391	0.404	0.378	0.403	0.432
20%	lin	0.706	0.654	0.702	0.754	0.784	0.724	0.779	0.837
	h_{cv}	0.628	0.58	0.626	0.672	3.002	2.441	3.122	3.417
	h_{wcv}	0.628	0.580	0.626	0.672	0.699	0.646	0.698	0.748
	h_{oracle}	0.622	0.575	0.620	0.667	0.682	0.630	0.679	0.731
25%	lin	1.087	1.011	1.080	1.156	1.214	1.134	1.210	1.287
	h_{bcv}	0.905	0.837	0.901	0.970	3.155	2.638	3.260	3.602
	h_{wcv}	0.905	0.837	0.901	0.970	1.009	0.936	1.004	1.076
	h_{oracle}	0.898	0.830	0.894	0.962	0.990	0.919	0.988	1.055

Table 4.4: Estimation errors according to $L(\widehat{\mu})$ for different noise levels and bandwidth values, with $d = 200$ time instants. Units are selected by simple random sampling without replacements (SRSWR) or stratified sampling.

4.5 Concluding remarks

We have studied in this paper the use of survey sampling methods for estimating a population mean temporal signal. This type of approach is extremely effective when data transmission or storage costs are important, in particular for large networks of distributed sensors. In view of noisy functional data, we have built a functional estimator by first smoothing the sampled curves and then setting up a Horvitz-Thompson estimator based on the smoothed curves. It has been shown that the estimator satisfies a CLT in the space of continuous functions and that its covariance can be estimated uniformly and consistently. These results have been exploited to show that by simulating Gaussian processes conditional on the estimated covariance, one obtains global confidence bands with asymptotic correct coverage. The problem of bandwidth selection, which is particularly difficult in the survey sampling context, has been addressed. We have devised a weighted cross-validation method that aims at mimicking an oracle estimator. This method has displayed very good performances in our numerical study; however, its rigorous theoretical study of remains to be done. Our numerical study has also revealed that in comparison to SRSWR, unequal probability sampling (e.g. stratified sampling) yields far superior performances and that when the noise level in the data is moderate to high, incorporating a smoothing step in the estimation procedure greatly enhances the accuracy in comparison to interpolation. Furthermore, we have seen that even when the noise level is low, smoothing can be highly beneficial for build global confidence bands. Indeed, smoothing the data leads

		SRSWR				Stratified sampling			
δ	h	Mean	1Q	Median	3Q	Mean	1Q	Median	3Q
5%	lin	0.044	0.042	0.044	0.047	0.049	0.047	0.049	0.051
	h_{cv}	0.040	0.038	0.040	0.042	2.231	1.612	1.917	2.806
	h_{wcv}	0.040	0.038	0.040	0.042	0.052	0.049	0.052	0.055
	h_{oracle}	0.040	0.038	0.040	0.042	0.044	0.041	0.044	0.046
10%	lin	0.175	0.166	0.175	0.184	0.196	0.186	0.196	0.205
	h_{cv}	0.128	0.120	0.127	0.135	2.258	1.648	1.969	2.844
	h_{wcv}	0.128	0.120	0.127	0.135	0.145	0.135	0.144	0.154
	h_{oracle}	0.127	0.119	0.127	0.134	0.138	0.13	0.137	0.146
15%	lin	0.397	0.377	0.397	0.416	0.444	0.420	0.445	0.466
	h_{cv}	0.233	0.217	0.232	0.247	2.293	1.682	1.991	2.901
	h_{wcv}	0.233	0.217	0.232	0.247	0.257	0.238	0.255	0.272
	h_{oracle}	0.231	0.215	0.230	0.246	0.250	0.234	0.250	0.266
20%	lin	0.708	0.672	0.706	0.744	0.79	0.749	0.791	0.829
	h_{cv}	0.351	0.327	0.350	0.374	2.442	1.763	2.152	3.107
	h_{wcv}	0.351	0.327	0.350	0.374	0.388	0.359	0.384	0.413
	h_{oracle}	0.349	0.326	0.348	0.373	0.381	0.355	0.380	0.406
25%	lin	1.089	1.030	1.087	1.142	1.219	1.155	1.212	1.280
	h_{cv}	0.498	0.462	0.495	0.535	2.591	1.932	2.344	3.254
	h_{wcv}	0.498	0.462	0.495	0.535	0.552	0.509	0.549	0.594
	h_{oracle}	0.497	0.460	0.494	0.533	0.547	0.505	0.545	0.586

Table 4.5: Estimation errors according to $L(\widehat{\mu})$ for different noise levels and bandwidth values, with $d = 400$ time instants. Units are selected by simple random sampling without replacements (SRSWR) or stratified sampling.

to estimators that have higher temporal correlation, which in turn makes the confidence bands narrower and more stable. Our method for confidence bands is simple and quick to implement. It gives satisfactory coverage (a little conservative) when the bandwidth is chosen correctly, e.g. with our weighted cross-validation method. Such confidence bands can find a variety of applications in statistical testing. They can be used to compare mean functions in different sub-populations, or to test for a parametric shape or for periodicity, among others. Examples of applications can be found in Degras (2010).

This work also raises some questions which deserve further investigation. A straightforward extension could be to relax the normality assumption made on the measurement errors. It is possible to consider more general error distributions under additional assumptions on the moments and much longer proofs. In another direction, it would be worthwhile to see whether our methodology can be extended to build confidence bands for other functional parameters such as population quantile or covariance functions. Also, as mentioned earlier, the weighted cross-validation proposed in this work seems a promising candidate for automatic bandwidth selection. However it is for now only based on heuristic arguments and its theoretical underpinning should be investigated.

Finally, it is well known that taking account of auxiliary information, which can be made available for all the units of the population at a low cost, can lead to substantial improvements with model assisted estimators (Särndal *et al.* 1992). In a functional context, an interesting strategy consists in first reducing the dimension through a functional principal components analysis shaped for the sampling framework (Cardot *et. al* 2010a)

		SRSWR					Stratified sampling				
δ	h	$1-\widehat{\alpha}$	Mean	1Q	Median	3Q	$1-\widehat{\alpha}$	Mean	1Q	Median	3Q
5%	lin	97.2	10.91	10.74	10.90	11.07	98.1	5.946	5.868	5.946	6.019
	h_{cv}	97.3	10.89	10.73	10.89	11.06	47.5	5.681	5.600	5.680	5.760
	h_{wcv}	97.3	10.89	10.73	10.89	11.06	97.5	5.918	5.840	5.913	6.000
	h_{oracle}	97.2	10.9	10.72	10.90	11.07	98.0	5.941	5.862	5.942	6.018
	$\widehat{\mu}_s$	97.3	10.54	10.36	10.54	10.70	98.2	5.593	5.513	5.596	5.671
10%	lin	98.1	11.43	11.25	11.42	11.60	97.1	6.455	6.374	6.458	6.531
	h_{cv}	98.0	11.38	11.20	11.37	11.55	50.2	5.903	5.819	5.902	5.980
	h_{wcv}	98.0	11.38	11.20	11.37	11.55	96.4	6.358	6.277	6.355	6.433
	h_{oracle}	98.1	11.39	11.22	11.39	11.56	96.7	6.414	6.335	6.416	6.496
	$\widehat{\mu}_s$	97.7	10.54	10.36	10.53	10.72	97.1	5.597	5.515	5.598	5.671
15%	lin	98.0	12.03	11.84	12.03	12.19	98.4	7.024	6.942	7.023	7.104
	h_{cv}	97.8	11.88	11.71	11.89	12.04	51.4	6.161	6.066	6.159	6.252
	h_{wcv}	97.8	11.88	11.71	11.89	12.04	98.2	6.804	6.720	6.799	6.891
	h_{oracle}	97.8	11.89	11.71	11.89	12.06	98.4	6.876	6.782	6.878	6.964
	$\widehat{\mu}_s$	97.3	10.56	10.38	10.56	10.72	98.2	5.598	5.519	5.594	5.679
20%	lin	98.5	12.61	12.44	12.62	12.80	97.5	7.631	7.537	7.628	7.724
	h_{cv}	98.4	12.29	12.12	12.29	12.45	54.0	6.418	6.316	6.410	6.509
	h_{wcv}	98.4	12.29	12.12	12.29	12.45	97.0	7.198	7.105	7.195	7.286
	h_{oracle}	98.3	12.32	12.14	12.31	12.50	97.3	7.306	7.212	7.305	7.393
	$\widehat{\mu}_s$	98.2	10.54	10.38	10.55	10.70	97.0	5.595	5.512	5.597	5.676
25%	lin	97.7	13.23	13.06	13.22	13.41	98.3	8.270	8.185	8.269	8.357
	h_{cv}	97.2	12.66	12.49	12.65	12.83	64.7	6.704	6.603	6.691	6.790
	h_{wcv}	97.2	12.66	12.49	12.65	12.83	97.3	7.563	7.479	7.564	7.645
	h_{oracle}	97.3	12.70	12.50	12.70	12.87	97.5	7.683	7.575	7.678	7.788
	$\widehat{\mu}_s$	97.0	10.53	10.37	10.52	10.70	97.7	5.589	5.514	5.586	5.663

Table 4.6: Empirical covering levels $1-\widehat{\alpha}$ and confidence band areas for different noise levels and bandwidth values, with $d = 200$ time instants. Units are selected by simple random sampling without replacements (SRSWR) or stratified sampling.

and then consider semi parametric models relating the principal components scores to the auxiliary variables (Cardot *et al.* 2010b). It is still possible to get consistent estimators of the covariance function of the limit process but further investigations are needed to prove the functional asymptotic normality and deduce that Gaussian simulations based approaches still lead to accurate confidence bands.

Acknowledgement. Etienne Josserand thanks the *Conseil Régional de Bourgogne, France* for its financial support (FABER PhD grant).

Appendix

Throughout the proofs we use the letter C to denote a generic constant whose value may vary from place to place. This constant does not depend on N nor on the arguments $s, t \in [0, T]$.

Proof of Theorem 4.3.1. We first decompose the difference between the estimator $\widehat{\mu}_N(t)$ and its target $\mu_N(t)$ as the sum of two stochastic components, one pertaining to the

		SRSWR					Stratified sampling				
δ	h	$1-\widehat{\alpha}$	Mean	1Q	Median	3Q	$1-\widehat{\alpha}$	Mean	1Q	Median	3Q
5%	lin	97.4	10.97	10.81	10.97	11.15	97.9	6.027	5.948	6.024	6.106
	h_{cv}	97.5	10.90	10.73	10.91	11.06	48.4	5.640	5.566	5.634	5.717
	h_{wcv}	97.5	10.90	10.73	10.91	11.06	97.6	5.894	5.816	5.889	5.971
	h_{oracle}	97.4	10.92	10.75	10.92	11.09	97.6	5.964	5.883	5.959	6.040
	$\widehat{\mu}_s$	97.3	10.54	10.38	10.54	10.70	97.8	5.597	5.519	5.591	5.675
10%	lin	97.8	11.58	11.41	11.57	11.76	97.8	6.589	6.504	6.590	6.669
	h_{cv}	97.6	11.23	11.06	11.22	11.40	49.5	5.788	5.705	5.786	5.864
	h_{wcv}	97.6	11.23	11.06	11.22	11.40	97.1	6.173	6.090	6.173	6.254
	h_{oracle}	97.7	11.25	11.08	11.24	11.41	97.3	6.257	6.180	6.256	6.333
	$\widehat{\mu}_s$	97.5	10.55	10.39	10.54	10.71	97.5	5.592	5.517	5.592	5.668
15%	lin	97.4	12.23	12.05	12.23	12.40	98.0	7.218	7.130	7.212	7.307
	h_{cv}	96.8	11.50	11.33	11.50	11.67	52.7	5.965	5.880	5.962	6.048
	h_{wcv}	96.8	11.50	11.33	11.50	11.67	97.3	6.453	6.366	6.447	6.533
	h_{oracle}	96.8	11.51	11.34	11.51	11.67	97.7	6.503	6.417	6.497	6.589
	$\widehat{\mu}_s$	96.7	10.55	10.38	10.56	10.72	97.7	5.590	5.510	5.588	5.668
20%	lin	98.2	12.90	12.72	12.91	13.08	98.2	7.891	7.805	7.892	7.972
	h_{cv}	97.7	11.80	11.63	11.80	11.96	55.1	6.153	6.071	6.154	6.235
	h_{wcv}	97.7	11.80	11.63	11.80	11.96	97.4	6.764	6.673	6.759	6.841
	h_{oracle}	97.7	11.79	11.62	11.79	11.96	97.6	6.785	6.700	6.783	6.868
	$\widehat{\mu}_s$	97.9	10.56	10.39	10.56	10.72	97.5	5.598	5.518	5.598	5.672
25	lin	98.0	13.58	13.40	13.58	13.75	98.3	8.587	8.491	8.588	8.676
	h_{cv}	97.5	12.11	11.95	12.10	12.28	58.1	6.344	6.244	6.343	6.437
	h_{wcv}	97.5	12.11	11.95	12.10	12.28	97.6	7.088	6.998	7.081	7.172
	h_{oracle}	97.5	12.12	11.94	12.12	12.29	97.8	7.101	7.011	7.099	7.188
	$\widehat{\mu}_s$	97.4	10.56	10.39	10.55	10.73	97.6	5.592	5.509	5.590	5.668

Table 4.7: Empirical covering levels $1-\widehat{\alpha}$ and confidence band areas for different noise levels and bandwidth values, with $d = 400$ time instants. Units are selected by simple random sampling without replacements (SRSWR) or stratified sampling.

sampling variability and the other to the measurement errors, and of a deterministic bias component:

$$\widehat{\mu}_N(t) - \mu_N(t) = \frac{1}{N}\sum_{k\in U}\left(\frac{I_k}{\pi_k} - 1\right)\tilde{X}_k(t) + \frac{1}{N}\sum_{k\in U}\frac{I_k}{\pi_k}\tilde{\epsilon}_k(t) + \frac{1}{N}\sum_{k}\left(\tilde{X}_k(t) - X_k(t)\right) \quad (4.22)$$

where $\tilde{X}_k(t)$ and $\tilde{\epsilon}_k(t)$ are defined in (4.9).

Bias term.
To study the bias term $N^{-1}\sum_k\left(\tilde{X}_k(t) - X_k(t)\right) = E(\widehat{\mu}_N(t)) - \mu_N(t)$ in (4.22), it suffices to use classical results on local linear smoothing (e.g. Tsybakov (2009), Proposition 1.13) together with the Hölder continuity (A2) of the X_k to see that

$$\sup_{t\in[0,T]}\left|\frac{1}{N}\sum_{k}\left(\tilde{X}_k(t) - X_k(t)\right)\right| \leq \frac{1}{N}\sum_{k}\sup_{t\in[0,T]}\left|\tilde{X}_k(t) - X_k(t)\right| \leq Ch^{\beta}. \quad (4.23)$$

Hence, for the bias to be negligible in the normalized estimator, it is necessary that the bandwidth satisfy $\sqrt{N}h^{\beta} \to 0$ as $N \to \infty$.

Error term.
We now turn to the measurement error term in (4.22), which can be seen as a sequence of random functions. We first show that this sequence goes pointwise to zero in mean square (a fortiori in probability) at a rate $(Ndh)^{-1}$. We then establish its tightness in $C([0,T])$, when premultiplied by $\sqrt{N}$, to prove the uniformity of the convergence over $[0,T]$.

Writing the vector of local linear weights at point t as follows

$$W(t) = (W_1(t), \ldots, W_d(t))'$$

and using the i.i.d assumption (A4) on the $(\epsilon_{k1}, \ldots, \epsilon_{kd})', k \in U_N$, we first obtain that

$$\begin{aligned} E\left(\frac{1}{N}\sum_{k\in U}\frac{I_k}{\pi_k}\tilde{\epsilon}_k(t)\right)^2 &= \frac{1}{N^2}\sum_{k\in U}\frac{1}{\pi_k}E\left(\tilde{\epsilon}_k(t)\right)^2 \\ &= \frac{1}{N^2}\sum_{k\in U}\frac{1}{\pi_k}W(t)'\mathbf{V}_N W(t). \end{aligned}$$

Then, considering the facts that $\min_k \pi_k > c$ by (A2), $\|\mathbf{V}_N\|$ is uniformly bounded in N by (A4), and exploiting a classical bound on the weights of the local linear smoother (e.g. Tsybakov (2009), Lemma 1.3), we deduce that

$$\begin{aligned} E\left(\frac{1}{N}\sum_{k\in U}\frac{I_k}{\pi_k}\tilde{\epsilon}_k(t)\right)^2 &\leq \frac{N}{(\min \pi_k)N^2}\left\|W(t)\right\|^2\left\|\mathbf{V}_N\right\| \\ &\leq \frac{C}{Ndh}\,. \end{aligned} \tag{4.24}$$

We can now prove the tightness of the sequence of processes $(N^{-1/2}\sum_k (I_k/\pi_k)\,\tilde{\epsilon}_k)$. Let us define the associated pseudo-metric

$$d_\epsilon^2(s,t) = \mathbb{E}\left(\frac{1}{\sqrt{N}}\sum_{k\in U}\frac{I_k}{\pi_k}\left(\tilde{\epsilon}_k(s) - \tilde{\epsilon}_k(t)\right)\right)^2.$$

We use the following maximal inequality holding for sub-Gaussian processes (van der Vaart and Wellner (2000), Corollary 2.2.8):

$$E\left(\sup_{t\in[0,T]}\left|\frac{1}{\sqrt{N}}\sum_{k\in U}\frac{I_k}{\pi_k}\tilde{\epsilon}_k(t)\right|\right) \leq E\left(\left|\frac{1}{\sqrt{N}}\sum_{k\in U}\frac{I_k}{\pi_k}\tilde{\epsilon}_k(t_0)\right|\right) + K\int_0^\infty \sqrt{\log N(x,d_\epsilon)}dx\,, \tag{4.25}$$

where t_0 is an arbitrary point in $[0,T]$ and the covering number $N(x,d_\epsilon)$ is the minimal number of d_ϵ-balls of radius $x>0$ needed to cover $[0,T]$. Note the equivalence of working with packing or covering numbers in maximal inequalities, see *ibid* p. 98. Also note that the sub-Gaussian nature of the smoothed error process $N^{-1/2}\sum_{k\in U}(I_k/\pi_k)\,\tilde{\epsilon}_k$ stems from the i.i.d. multivariate normality of the random vectors $(\epsilon_{k1}, \ldots, \epsilon_{kd})'$ and the boundedness of the I_k for $k \in U_N$.

By the arguments used in (4.24) and an elementary bound on the increments of the

weight function vector W (see e.g. Lemma 1 in Degras (2010)), one obtains that

$$\begin{aligned} d_\epsilon^2(s,t) &= \frac{1}{N}\sum_{k\in U}\frac{1}{\pi_k}E\left(\tilde{\epsilon}_k(s)-\tilde{\epsilon}_k(t)\right)^2 \\ &\leq \frac{1}{\min \pi_k}\|W(s)-W(t)\|^2\,\|\mathbf{V}_N\| \\ &\leq \frac{C}{dh}\left(\frac{|s-t|^2}{h^2}\wedge 1\right). \end{aligned} \tag{4.26}$$

It follows that the covering numbers satisfy

$$\begin{cases} N(x,d_\epsilon)=1, & \text{if } \frac{C}{dh}\leq x^2, \\ N(x,d_\epsilon)\leq \frac{\sqrt{C}}{h\sqrt{dh}x}, & \text{if } \frac{C}{dh}>x^2. \end{cases}$$

Plugging this bound and the pointwise convergence (4.24) in the maximal inequality (4.25), we get after a simple integral calculation (see Eq. (17) in Degras (2010) for details) that

$$E\left(\sup_{t\in[0,T]}\left|\frac{1}{\sqrt{N}}\sum_{k\in U}\frac{I_k}{\pi_k}\tilde{\epsilon}_k(t)\right|\right)\leq \frac{C}{dh}+C\sqrt{\frac{|\log(h)|}{dh}}\,. \tag{4.27}$$

Thanks to Markov's inequality, the previous bound guarantees the uniform convergence in probability of $N^{-1/2}\sum_{k\in U}(I_k/\pi_k)\tilde{\epsilon}_k$ to zero, provided that $|\log(h)|/(dh)\to 0$ as $N\to\infty$. The last condition is equivalent to $\log(d)/(dh)\to 0$ by the fact that $dh\to\infty$ and by the properties of the logarithm.

Main term: sampling variability.
Finally, we look at the process $N^{-1}\sum_{k\in U}\left(I_k/\pi_k-1\right)\tilde{X}_k$ in (4.22), which is asymptotically normal in $C([0,T])$ as we shall see. We first establish the finite-dimensional asymptotic normality of this process normalized by $\sqrt{N}$, after which we will prove its tightness thanks to a maximal inequality.

Let us start by verifying that the limit covariance function of the process is indeed the function γ defined in Section 4.3.1. The finite-sample covariance function expresses as

$$\begin{aligned} E\left\{\left(\frac{1}{\sqrt{N}}\sum_{k\in U}\left(\frac{I_k}{\pi_k}-1\right)\tilde{X}_k(s)\right)\,\left(\frac{1}{\sqrt{N}}\sum_{l\in U}\left(\frac{I_l}{\pi_l}-1\right)\tilde{X}_l(t)\right)\right\} & \\ = \frac{1}{N}\sum_{k,l\in U}\frac{\Delta_{kl}}{\pi_k\pi_l}\tilde{X}_k(s)\tilde{X}_l(t) & \\ = \frac{1}{N}\sum_{k,l\in U}\frac{\Delta_{kl}}{\pi_k\pi_l}X_k(s)X_l(t)+\mathcal{O}(h^\beta) & \\ = \gamma(s,t)+o(1)+\mathcal{O}(h^\beta). & \end{aligned} \tag{4.28}$$

To derive the previous relation we have used the facts that

$$\max_{k,l\in U}\sup_{s,t\in[0,T]}\left|\tilde{X}_k(s)\tilde{X}_l(t)-X_k(s)X_l(t)\right|\leq Ch^\beta$$

by (4.23) and the uniform boundedness of the X_k arising from (A2) and that, by (A1),

$$\begin{aligned}\frac{1}{N}\sum_{k,l\in U}\frac{|\Delta_{kl}|}{\pi_k\pi_l} &= \frac{1}{N}\sum_{k\neq l}\frac{|\Delta_{kl}|}{\pi_k\pi_l}+\frac{1}{N}\sum_k\frac{\Delta_{kk}}{\pi_k^2}\\ &\leq \frac{1}{N}\frac{N(N-1)}{2}\frac{\max_{k,l}(n|\Delta_{kl}|)}{n}+\frac{1}{N}\sum_k\frac{1-\pi_k}{\pi_k}\leq C. \end{aligned} \tag{4.29}$$

We now check the finite-dimensional convergence of $N^{-1/2}\sum_{k\in U}(I_k/\pi_k-1)\,\tilde{X}_k$ to a centered Gaussian process with covariance γ. In light of the Cramer-Wold theorem, this convergence is easily shown with characteristic functions and appears as a straightforward consequence of(A5). It suffices for us to check that the uniform boundedness of the trajectories X_k derived from (A2) is preserved by local linear smoothing, so that the $\tilde{X}_k$ are uniformly bounded as well.

It remains to establish the tightness of the previous sequence of processes so as to obtain its asymptotic normality in $C([0,T])$. To that intent we use the maximal inequality of the Corollary 2.2.5 in van der Vaart and Wellner (2000). With the notations of this result, we consider the pseudo-metric $d^2_{\tilde{X}}(s,t)=E\{N^{-1/2}\sum_{k\in U}(I_k/\pi_k-1)(\tilde{X}_k(s)-\tilde{X}_k(t))\}^2$ and the function $\psi(t)=t^2$ for the Orlicz norm. We get the following bound for the second moment of the maximal increment:

$$\begin{aligned}E\left\{\sup_{d_{\tilde{X}}(s,t)\leq\delta}\left|\frac{1}{\sqrt{N}}\sum_{k\in U}\left(\frac{I_k}{\pi_k}-1\right)\left(\tilde{X}_k(s)-\tilde{X}_k(t)\right)\right|\right\}^2 \\ \leq C\left(\int_0^\eta\psi^{-1}(N(x,d_{\tilde{X}}))dx+\delta\psi^{-1}(N^2(\eta,d_{\tilde{X}}))\right)^2\end{aligned} \tag{4.30}$$

for any arbitrary constants $\eta,\delta>0$. Observe that the maximal inequality (4.30) is weaker than (4.25) where an additional assumption of sub-Gaussianity is made (no log factor in the integral above). Employing again the arguments of (4.28), we see that

$$\begin{aligned}d^2_{\tilde{X}}(s,t) &= \frac{1}{N}\sum_{k,l}\frac{\Delta_{kl}}{\pi_k\pi_l}\left(\tilde{X}_k(s)-\tilde{X}_k(t)\right)\left(\tilde{X}_l(s)-\tilde{X}_l(t)\right)\\ &\leq \frac{C}{N}\frac{N(N-1)}{2n}|s-t|^{2\beta}+\frac{C}{N}N|s-t|^{2\beta}\\ &\leq C|s-t|^{2\beta}.\end{aligned} \tag{4.31}$$

It follows that the covering number satisfies $N(x,d_{\tilde{X}})\leq Cx^{-1/\beta}$ and that the integral in (4.30) is smaller than $C\int_0^\eta x^{-0.5/\beta}dx=C\eta^{1-0.5/\beta}$, which can be made arbitrarily small since $\beta>0.5$. Once η is fixed, δ can be adjusted to make the other term in the right-handside of (4.30) arbitrarily small as well. With Markov's inequality, we deduce that the sequence $(N^{-1/2}\sum_{k\in U}(I_k/\pi_k-1)\,\tilde{X}_k)_{N\geq 1}$ is asymptotically $d_{\tilde{X}}$-equicontinuous in probability (with the terminology of van der Vaart and Wellner (2000)), which guarantees its tightness in $C([0,T])$. □

Proof of Theorem 4.3.2.

Mean square convergence.
We first decompose the distance between $\widehat{\gamma}_N(s,t)$ and its target $\gamma_N(s,t)$ as follows:

$$\begin{aligned}\widehat{\gamma}_N(s,t) - \gamma_N(s,t) &= \frac{1}{N}\sum_{k,l\in U}\frac{\Delta_{kl}}{\pi_k\pi_l}\left(\frac{I_kI_l}{\pi_{kl}}-1\right)\tilde{X}_k(s)\tilde{X}_l(t)\\ &+\frac{2}{N}\sum_{k,l\in U}\frac{\Delta_{kl}}{\pi_k\pi_l}\frac{I_kI_l}{\pi_{kl}}\tilde{X}_k(s)\tilde{\epsilon}_l(t)\\ &+\frac{1}{N}\sum_{k,l\in U}\frac{\Delta_{kl}}{\pi_k\pi_l}\frac{I_kI_l}{\pi_{kl}}\tilde{\epsilon}_k(s)\tilde{\epsilon}_l(t)\\ &-\frac{1}{N}\sum_{k\in U}\frac{1}{\pi_k}E(\tilde{\epsilon}_k(s)\tilde{\epsilon}_k(t))\\ &:= A_{1,N}+A_{2,N}+A_{3,N}-A_{4,N}. \end{aligned} \tag{4.32}$$

To establish the mean square convergence of $(\widehat{\gamma}_N(s,t)-\gamma_N(s,t))$ to zero as $N\to\infty$, it is enough to show that $E(A_{i,N}^2)\to 0$ for $i=1,\dots,4$, by the Cauchy-Schwarz inequality.

Let us start with

$$E(A_{1,N}^2) = \frac{1}{N^2}\sum_{k,l}\sum_{k',l'}\frac{\Delta_{kl}\Delta_{k'l'}}{\pi_k\pi_l\pi_{k'}\pi_{l'}}\frac{\mathbb{E}\{(I_kI_l-\pi_{kl})(I_{k'}I_{l'}-\pi_{k'l'})\}}{\pi_{kl}\pi_{k'l'}}\widetilde{X}_k(s)\widetilde{X}_l(t)\widetilde{X}_{k'}(s)\widetilde{X}_{l'}(t). \tag{4.33}$$

It can be shown that this sum converges to zero by strictly following the proof of the Theorem 3 in Breidt and Opsomer (2000). The idea of the proof is to partition the set of indexes in (4.33) into (i) $k=l$ and $k'=l'$, (ii) $k=l$ and $k'\neq l'$ or vice-versa, (iii) $k\neq l$ and $k'\neq l'$, and study the related subsums. The convergence to zero is then handled with assumption (A1) (mostly) in case (i), with (A1)-(A6) in case (iii), and thanks to the previous results and Cauchy-Schwarz inequality in case (ii). More precisely, it holds that

$$\begin{aligned}E(A_{1,N}^2) &\le \frac{C\max_{k\neq l} n|\Delta_{kl}|}{(\min\pi_k)^4 n}+\frac{C}{(\min\pi_k)^3 N}\\ &+\left(\frac{C(\max_{k\neq l}n|\Delta_{kl}|)N}{(\min\pi_k)^2(\min_{k\neq l}\pi_{kl})n}\right)^2\max_{(k,l,k',l')\in D_{4,N}}|\mathbb{E}\{(I_kI_l-\pi_{kl})(I_{k'}I_{l'}-\pi_{k'l'})\}|.\end{aligned} \tag{4.34}$$

For the (slightly simpler) study of $E(A_{2,N}^2)$, we provide an explicit decomposition:

$$\begin{aligned}E(A_{2,N}^2) &= \frac{4}{N^2}\sum_{k,l}\sum_{k'}\frac{\Delta_{kl}\Delta_{k'l}}{\pi_k\pi_{k'}\pi_l^2}\tilde{X}_k(s)\tilde{X}_{k'}(t)\,E(\tilde{\epsilon}_l(s)\tilde{\epsilon}_l(t))\\ &= \frac{4}{N^2}\sum_{k\in U}\frac{\Delta_{kk}^2}{\pi_k^5}\tilde{X}_k(s)\tilde{X}_k(t)E(\tilde{\epsilon}_k(s)\tilde{\epsilon}_k(t))\\ &+\frac{4}{N^2}\sum_{k\neq k'}\frac{\Delta_{kk}\Delta_{kk'}}{\pi_k^4\pi_{k'}}\tilde{X}_k(s)\tilde{X}_{k'}(t)\,E(\tilde{\epsilon}_k(s)\tilde{\epsilon}_k(t))\\ &+\frac{4}{N^2}\sum_{k\neq l}\sum_{k':k'\neq l}\frac{\Delta_{kl}\Delta_{k'l}}{\pi_k\pi_{k'}\pi_l^2}\tilde{X}_k(s)\tilde{X}_{k'}(t)\,E(\tilde{\epsilon}_l(s)\tilde{\epsilon}_l(t)).\end{aligned} \tag{4.35}$$

Note that the expression of $E(A_{2,N}^2)$ as a quadruple sum over $k,l,k',l'\in U_N$ reduces to a triple sum since $E(\tilde{\epsilon}_l(s)\tilde{\epsilon}_{l'}(t))=0$ if $l\neq l'$ by (A4). With the bound $|E(\tilde{\epsilon}_k(s)\tilde{\epsilon}_k(t))|=$

$|W(s)'\mathbf{V}_N W(t)| \leq \|W(s)\|\|\mathbf{V}_N\|\|W(t)\| \leq C/(dh)$, it follows that

$$E(A_{2,N}^2) \leq \frac{CN}{N^2}\frac{\|\mathbf{V}_N\|}{dh} + \frac{CN^2}{N^2}\frac{\max_{k\neq k'} n|\Delta_{kk'}|}{n}\frac{\|\mathbf{V}_N\|}{dh} + \frac{CN^3}{N^2}\frac{(\max_{k\neq l} n|\Delta_{kl}|)^2}{n^2}\frac{\|\mathbf{V}_N\|}{dh} = \frac{C}{Ndh}. \tag{4.36}$$

To study the term $E(A_{3,N}^2)$, we start with the same partition of the quadruple sum as the one used with $E(A_{1,N}^2)$. Here, due to the independence assumption (A4) on the error vectors, the partition simplifies further into (i) $k = l$, $k' = l'$, $k \neq k'$, and (ii) $k = l = k' = l'$:

$$E(A_{3,N}^2) = \frac{1}{N^2}\sum_{k\neq k'} \frac{\Delta_{kk'}}{\pi_k\pi_{k'}}\frac{I_k I_{k'}}{\pi_{kk'}} E(\tilde{\epsilon}_k(s)\tilde{\epsilon}_k(t))\, E(\tilde{\epsilon}_{k'}(s)\tilde{\epsilon}_{k'}(t)) + \frac{1}{N^2}\sum_k \frac{\Delta_{kk}}{\pi_k^2}\frac{I_k}{\pi_{kk}} E(\tilde{\epsilon}_k^2(s)\tilde{\epsilon}_k^2(t)). \tag{4.37}$$

Forgoing the calculations already done before, we focus on the main task which for this term is to bound the quantity $E(\tilde{\epsilon}_k^2(s)\tilde{\epsilon}_k^2(t))$ (recall that $E(\tilde{\epsilon}_k(s)\tilde{\epsilon}_k(t)) \leq C/(dh)$ as seen before). We first note that $E(\tilde{\epsilon}_k^2(s)\tilde{\epsilon}_k^2(t)) \leq \{E(\tilde{\epsilon}_k^4(s))\}^{1/2}\{E(\tilde{\epsilon}_k^4(t))\}^{1/2}$. Writing $\boldsymbol{\epsilon} \sim N(0, \mathbf{V}_N)$, it holds that $E(\tilde{\epsilon}_k^4(t)) = E((W(t)'\boldsymbol{\epsilon})^4) = 3(W(t)'\mathbf{V}_N W(t))^2$ by the moment properties of the normal distribution. Plugging this expression in (4.37), we find that

$$E(A_{3,N}^2) \leq \frac{C}{(dh)^2} + \frac{C}{N(dh)^2}. \tag{4.38}$$

Finally, note that an expression very similar to the deterministic term $A_{4,N}(s,t)$ has been studied in (4.24). One easily concludes that $A_{4,N}(s,t)$ is dominated by $(dh)^{-1}$ uniformly in $s, t \in [0,T]$.

Tightness.
To prove the tightness of the sequence $(\widehat{\gamma}_N - \gamma_N)_{N\geq 1}$ in $C([0,T]^2)$, we study separately each term in the decomposition (4.32) and we call again to the maximal inequalities of van der Vaart and Wellner (2000).

For the first term $A_{1,N} = A_{1,N}(s,t)$, we consider the pseudo-metric d defined as the L^4-norm of the increments: $d_1^4((s,t),(s',t')) = E|A_{1,N}(s,t) - A_{1,N}(s',t')|^4$. (The need to use here the L^4-norm and not the usual L^2-norm is justified hereafter by a dimension argument.) With (A1)-(A2) and the approximation properties of local linear smoothers, one sees that

$$\left|\frac{1}{N}\sum_{k,l\in U}\frac{\Delta_{kl}}{\pi_k\pi_l}\left(\frac{I_k I_l}{\pi_{kl}} - 1\right)\left(\tilde{X}_k(s)\tilde{X}_l(t) - \tilde{X}_k(s')\tilde{X}_l(t')\right)\right| \leq C\left(|s-s'|^\beta + |t-t'|^\beta\right).$$

Hence $d_1(s,t) \leq C\left(|s-s'|^\beta + |t-t'|^\beta\right)$ and for all $x > 0$, the covering number $N(x, d_1)$ is no larger than the size of a two-dimensional square grid of mesh $x^{1/\beta}$, i.e. $N(x,d_1) \leq Cx^{-2/\beta}$. (Compare to the proof of Theorem 1 where, for the main term $N^{-1/2}\sum_k (I_k/\pi_k)\tilde{X}_k$, we

have $N(x, d_{\tilde{X}}) \leq Cx^{-1/\beta}$ because the index set $[0,T]$ is of dimension 1.) Using Theorem 2.2.4 of van der Vaart and Wellner (2000) with $\psi(t) = t^4$, it follows that for all $\eta, \delta > 0$,

$$E\left\{\sup_{d_1((s,t),(s',t'))\leq\delta} |A_{1,N}(s,t) - A_{1,N}(s',t')|^4\right\} \leq C\left(\int_0^\eta \psi^{-1}(N(x,d_1))dx + \delta\psi^{-1}(N^2(\eta,d_1))\right)^4$$
$$\leq C\left(\eta^{1-0.5/\beta} + \delta\eta^{-1/\beta}\right)^4.$$

The upper bound above can be made arbitrarily small by varying η first and δ next since $\beta > 0.5$. Hence, with Markov's inequality, we deduce that the processes $A_{1,N}$ are tight in $C([0,T]^2)$.

The bivariate processes $(A_{2,N})_{N\geq 1}$ are sub-Gaussian for the same reasons as the univariate processes $N^{-1/2}\sum_{k\in U}(I_k/\pi_k)\,\tilde{\epsilon}_k$ are in the proof of Theorem 1, namely the independence and multivariate normality of the error vectors $(\epsilon_{k1},\ldots,\epsilon_{kd})'$ and the boundedness of the sample membership indicators I_k for $k \in U_N$. Therefore, although the covering number $N(x,d_2)$ grows to $O(x^{-2/\beta})$ in dimension 2, with d_2 being the L^2-norm on $[0,T]^2$, this does not affect significantly the integral upper bound $\int_0^\infty \sqrt{\log(N(x,d_2)}dx$ in a maximal inequality like (4.25). As a consequence, one obtains the tightness of $(A_{2,N})$ in $C([0,T]^2)$.

To study the term $A_{3,N}(s,t)$ in (4.32), we start with the following bound:

$$\begin{aligned}
|A_{3,N}(s,t)| &\leq \frac{1}{N}\sum_{k,l}\frac{|\Delta_{kl}|}{\pi_k\pi_l}\frac{I_kI_l}{\pi_{kl}}\frac{\tilde{\epsilon}_k^2(s)+\tilde{\epsilon}_l^2(t)}{2}\\
&= \frac{1}{N}\sum_k\left(\sum_l\frac{|\Delta_{kl}|}{2\pi_l}\frac{I_l}{\pi_{kl}}\right)\frac{I_k}{\pi_k}\tilde{\epsilon}_k^2(s) + \frac{1}{N}\sum_l\left(\sum_k\frac{|\Delta_{kl}|}{2\pi_k}\frac{I_k}{\pi_{kl}}\right)\frac{I_l}{\pi_l}\tilde{\epsilon}_l^2(t)\\
&\leq \frac{C}{N}\sum_k\tilde{\epsilon}_k^2(s) + \frac{C}{N}\sum_l\tilde{\epsilon}_l^2(t).
\end{aligned}$$

The two-dimensional study is thus reduced to an easier one-dimensional problem.

To apply the Corollary 2.2.5 of van der Vaart and Wellner (2000), we consider the function $\psi(t) = t^m$ and the pseudo-metric $d_3^m(s,t) = E|N^{-1}\sum_k(\tilde{\epsilon}_k^2(s) - \tilde{\epsilon}_k^2(t))|^m$, where $m \geq 1$ is an arbitrary integer. We have that

$$E\left\{\sup_{s,t\in[0,T]}\left|\frac{1}{N}\sum_k\left(\tilde{\epsilon}_k^2(s)-\tilde{\epsilon}_k^2(t)\right)\right|^m\right\} \leq C\left(\int_0^{D_T}(N(x,d_3))^{1/m}dx\right)^m \tag{4.39}$$

where $D_T = \sup_{s,t\in[0,T]} d_3(s,t)$ is the diameter of $[0,T]$ for d_3. Using the classical inequality, $|\sum_{k=1}^n a_k|^m \leq n^{m-1}\sum_{k=1}^n |a_k|^m$, for $m > 1$ and arbitrary real number $a_1,\ldots,a_n$, we get, with the Cauchy-Schwarz inequality and the moment properties of Gaussian random

vectors, that

$$
\begin{aligned}
d_3^m(s,t) &\leq \frac{1}{N}\sum_k E\left|\tilde{\epsilon}_k^2(s)-\tilde{\epsilon}_k^2(t)\right|^m \\
&\leq \frac{1}{N}\sum_k \left\{E\left|\tilde{\epsilon}_k(s)-\tilde{\epsilon}_k(t)\right|^{2m}\right\}^{1/2}\left\{E\left|\tilde{\epsilon}_k(s)+\tilde{\epsilon}_k(t)\right|^{2m}\right\}^{1/2} \\
&\leq \frac{C_m}{N}\sum_k \|W(s)-W(t)\|_{\mathbf{V}_N}^m \|W(s)+W(t)\|_{\mathbf{V}_N}^m \\
&\leq \frac{C'_m}{(dh)^m}\left(\frac{|s-t|}{h}\wedge 1\right)^m, \qquad (4.40)
\end{aligned}
$$

where $\|\mathbf{x}\|_{\mathbf{V}_N}=(\mathbf{x}'\mathbf{V}_N\mathbf{x})^{1/2}$ and C_m and C'_m are constants that only depends on m.

We deduce from (4.40) that the diameter D_T is at most of order $1/(dh)$ and that for all $0<x\leq 1/(dh)$, the covering number $N(x,d_3)$ is of order $1/(xdh^2)$. Hence the integral bound in (4.39) is of order $\int_0^{1/(dh)}(dh^2x)^{-1/m}dx \leq C(dh^2)^{-1/m}(dh)^{(1-1/m)}) = C/(dh)^{1+1/m}$. Therefore, if $dh^{1+\alpha}\to\infty$ for some $\alpha>0$, the sequence $(N^{-1}\sum_k(\tilde{\epsilon}_k^2))_{N\geq 1}$ tends uniformly to zero in probability which concludes the study of the term $(A_{3,N})_{N\geq 1}$ and the proof. □

Proof of Theorem 4.3.3.

We show here the weak convergence of $(\widehat{G}_N)$ to G in $C([0,T])$ conditionally on $\widehat{\gamma}_N$. This convergence, together with the uniform convergence of $\widehat{\gamma}_N$ to γ presented in Theorem 4.3.2, is stronger than the result of Theorem 4.3.3 required to build simultaneous confidence bands.

First, the finite-dimensional convergence of $(\widehat{G}_N)$ to G conditionally on $\widehat{\gamma}_N$ is a trivial consequence of Theorem 4.3.2.

Second, we show the tightness of $(\widehat{G}_N)$ in $C([0,T])$ (conditionally on $\widehat{\gamma}_N$) similarly to the study of $(A_{3,N})$ in the proof of Theorem 4.3.2. We start by considering the random pseudo-metric $\hat{d}_\gamma^m(s,t)=\mathbb{E}[(\widehat{G}_N(s)-\widehat{G}_N(t))^m\,|\,\widehat{\gamma}_N]$, where $m\geq 1$ is an arbitrary integer. By the moment properties of Gaussian random variables, it holds that

$$
\begin{aligned}
\hat{d}_\gamma^m(s,t) &= C_m\left[\frac{1}{N}\sum_{k,l\in U}\frac{\Delta_{kl}}{\pi_{kl}}\frac{I_kI_l}{\pi_k\pi_l}\big(\widehat{X}_k(s)-\widehat{X}_k(t)\big)\big(\widehat{X}_l(s)-\widehat{X}_l(t)\big)\right]^{m/2} \\
&\leq C_m\left[\frac{1}{N}\sum_{k,l\in U}\frac{|\Delta_{kl}|}{\pi_{kl}}\frac{I_kI_l}{\pi_k\pi_l}\big(\widehat{X}_k(s)-\widehat{X}_k(t)\big)^2\right]^{m/2} \\
&\leq C_m\left[\frac{2}{N}\sum_{k,l\in U}\frac{|\Delta_{kl}|}{\pi_{kl}}\frac{I_kI_l}{\pi_k\pi_l}\big(\tilde{X}_k(s)-\tilde{X}_k(t)\big)^2+\frac{2}{N}\sum_{k,l\in U}\frac{|\Delta_{kl}|}{\pi_{kl}}\frac{I_kI_l}{\pi_k\pi_l}\left(\tilde{\epsilon}_k(s)-\tilde{\epsilon}_k(t)\right)^2\right]^{m/2} \\
&\leq \frac{C_m}{2}\left[\frac{1}{N}\sum_k\big(\tilde{X}_k(s)-\tilde{X}_k(t)\big)^2\right]^{m/2}+\frac{C_m}{2}\left[\frac{1}{N}\sum_k\left(\tilde{\epsilon}_k(s)-\tilde{\epsilon}_k(t)\right)^2\right]^{m/2}. \qquad (4.41)
\end{aligned}
$$

Clearly, the first sum in the right-handside of (4.41) is dominated by $|s-t|^{m\beta}$ thanks to (A2) and the approximation properties of local linear smoothers. The second sum can be

viewed as a random quadratic form. Introducing the square root $\mathbf{V}_N^{1/2}$ of $\mathbf{V}_N$, we note that $\boldsymbol{\epsilon}_k = \mathbf{V}_N^{1/2}\boldsymbol{Z}_k$, with equality in distribution, for $k = 1, \ldots, N$, where the $\boldsymbol{Z}_k$ are i.i.d centered d-dimensional Gaussian vectors with identity covariance matrix.

Thus,

$$\begin{aligned}\frac{1}{N}\sum_k (\tilde{\epsilon}_k(s) - \tilde{\epsilon}_k(t))^2 &= (W(s) - W(t))' \left(\frac{1}{N}\sum_k \boldsymbol{\epsilon}_k\boldsymbol{\epsilon}_k'\right)(W(s) - W(t)) \\ &\leq \|W(s) - W(t)\|^2 \left\|\frac{1}{N}\sum_k \boldsymbol{\epsilon}_k\boldsymbol{\epsilon}_k'\right\| \\ &\leq \|W(s) - W(t)\|^2 \|\mathbf{V}_N\| \left\|\frac{1}{N}\sum_k \mathbf{Z}_k\mathbf{Z}_k'\right\| \end{aligned} \tag{4.42}$$

Now, the vector norm $\|W(s) - W(t)\|^2$ has already been studied in (4.26) and the sequence $(\|\mathbf{V}_N\|)$ is bounded by (A4). The remaining matrix norm in (4.42) is smaller than the largest eigenvalue, up to a factor N^{-1}, of a d-variate Wishart matrix with N degrees of freedom. By (A3) it holds that $d = o(N/\log\log N)$ and one can apply Theorem 3.1 in Fey *et al.* (2008), which states that for any fixed $\alpha \geq 1$,

$$\lim_{N\to\infty} -\frac{1}{N}\log\mathbb{P}\left(\left\|\frac{1}{N}\sum_k \mathbf{Z}_k\mathbf{Z}_k'\right\| \geq \alpha\right) = \frac{1}{2}(\alpha - 1 - \log\alpha)\,. \tag{4.43}$$

A immediate consequence of (4.43) is that $\left\|\frac{1}{N}\sum_k \mathbf{Z}_k\mathbf{Z}_k'\right\|$ remains almost surely bounded as $N \to \infty$. Note that the same result holds if instead of (A3), (d/N) remains bounded away from zero and infinity, thanks to the pioneer work of Geman (1980) on the norm of random matrices. Thus, there exists a deterministic constant $C \in (0, \infty)$ such that

$$\hat{d}_\gamma^m(s,t) \leq C|s-t|^{m\beta} + \frac{C}{(dh)^{m/2}}\left(\frac{|s-t|}{h} \wedge 1\right)^m \tag{4.44}$$

for all $s, t \in [0, T]$, with probability tending to 1 as $N \to \infty$. Similarly to the previous entropy calculations, one can show that there exists a constant $C \in (0, \infty)$ such that $N(x, \hat{d}_\gamma) \leq C(x^{-1/\beta} + (dh^3)^{-1/2}x^{-1})$ for all $x \leq (dh)^{-1}$ with probability tending to 1 as $N \to \infty$. Applying the maximal inequality of van der Vaart and Wellner (2000) (Th. 2.2.4) to the conditional increments of $\widehat{G}_N$, with $\phi(t) = t^m$ (usual L^m-norm), one finds a covering integral $\int_0^{1/(ph)}(N(x, \hat{d}_\gamma))^{1/2}dx$ of the order of $(dh)^{1/(m\beta)-1} + (dh^3)^{-1/(2m)}(dh)^{1/m-1}$. Hence the covering integral tends to zero in probability, provided that $h \to 0$ and $dh^{\frac{1+1/(2m)}{1-1/(2m)}} \to \infty$ as $N \to \infty$. Obvisouly, the latter condition on h holds for some integer $m \geq 1$ if $dh^{1+\alpha} \to \infty$ for some real $\alpha > 0$. Under this condition, the sequence $(\widehat{G}_N)$ is tight in $C([0, T])$ and therefore converges to G. □

Acknowledgements. Etienne Josserand thanks the Conseil Régional de Bourgogne for its financial support (Faber PhD grant).

Bibliography

Adler, R. J. and Taylor, J. (2007). *Random Fields and Geometry.* Springer, New York.

Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics*, **12**, 470-482.

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, **28**, 1026-1053.

Bunea, F., Ivanescu, A. and M. Wegkamp (2011). Adaptive inference for the mean of a Gaussian process in functional data. *J. Roy. Statist. Soc. Ser. B*, to appear.

Cardot, H., Chaouch, M., Goga, C. and C. Labruère (2010a). Properties of Design-Based Functional Principal Components Analysis, *J. Statist. Planning and Inference*, **140**, 75-91.

Cardot, H., Dessertaine, A., Josserand, E. (2010b). Semiparametric models with functional responses in survey sampling setting : model assisted estimation of electricity consumption curves. *Compstat 2010*, Eds Lechevallier, Y. and Saporta, G. Physica-Verlag, Springer, 411-420.

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.

Chiky, R. and Hébrail, G. (2008). Summarizing distributed data streams for storage in data warehouses. In DaWaK 2008, I-Y. Song, J. Eder and T. M. Nguyen, Eds. *Lecture Notes in Computer Science*, Springer, 65-74.

Claeskens, G. and van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Ann. Statist.*, **31**, 18521884.

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.

Degras, D. (2009). Nonparametric estimation of a trend based upon sampled continuous processes. *C. R. Math. Acad. Sci. Paris*, **347**, 191-194.

Degras, D. (2010). Simultaneous confidence bands for nonparametric regression with functional data. Accepted for publication at *Statistica Sinica*. http://arxiv.org/abs/0908.1980

Erdös, P. and Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hungar. Acad. Sci.*, **4**, 49-61.

Eubank, R.L. and Speckman P.L. (1993). Confidence Bands in Nonparametric Regression. *J. Amer. Statist. Assoc.*, **88**, 1287-1301.

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

Fey, A., van der Hofstad, R. and Klok, M. (2008). Large deviations for eigenvalues of sample covariance matrices, with applications to mobile communication systems. *Adv. in Appl. Probab.* **40**, 1048-1071.

Fuller, W.A. (2009). *Sampling Statistics.* John Wiley and Sons.

Geman, S. (1980). A Limit Theorem for the Norm of Random Matrices. *Ann. Probab.*, **8**, 252-261

Hart, J. D. and Wehrly, T. E. (1993). Consistency of cross-validation when the data are curves. *Stoch. Proces. Applic.*, **45**, 351361.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *J. Am. Statist. Ass.*, **77**, 89-96.

Krivobokova, T., Kneib, T. and G. Claeskens (2010). Simultaneous confidence bands for penalized spline estimators. *J. Am. Statist. Ass.*, **105**, 852-863.

Landau, H. and Shepp, L.A. (1970), On the supremum of a Gaussian process. *Sankhyã*, **32**, 369-378

Mas A. (2007). Testing for the mean of random curves: a penalization approach. *Statistical Inference for Stochastic Processes,* **10**, 147-163

Opsomer, J. D. and Miller, C. P. (2005). Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. *J. Nonparametric Statistics,* **17**, 593-611.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B*, **53**, 233-243.

Robinson, P. M. and Särndal, C. E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya : The Indian Journal of Statistics*, **45**, 240-248.

Sun, J. and Loader, C.R. (1994). Simultaneous confidence bands for linear regression and smoothing. *Annals of Statistics*, **22**, 1328-1345.

Särndal, C. E., Swensson, B. and J. Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag.

Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation.* Springer, New York.

van der Vaart, A. D. and Wellner, J. A. (2000). *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer-Verlag, New York.

Zhang, J. T. and Chen, J. (2007). Statistical inferences for functional data. *Annals of Statistics*, **35**, 1052-1079.

Chapitre 5

Semiparametric models with functional response in a survey sampling setting : model assisted estimation of electricity consumption curves *

Résumé : Ce travail adopte un point de vue échantillonnage pour estimer la courbe moyenne de grandes bases de données fonctionnelles. Quand les capacités de stockage sont limitées, la sélection, avec des techniques d'enquête d'une petite fraction des observations est une alternative intéressante aux techniques de compression du signal. Nous proposons ici de tenir compte de l'information auxiliaire réelle ou multivariée disponible à un faible coût sur l'ensemble de la population, avec une approche modèle assisté semi-paramétrique, afin d'améliorer la précision des estimateurs de Horvitz-Thompson de la courbe moyenne. Nous estimons d'abord les composantes principales fonctionnelles avec un point de vue *design-based* dans le but de réduire la dimension des signaux et d'ensuite proposer des modèles semi-paramétriques pour obtenir des estimations des courbes qui ne sont pas observées. Cette technique s'est révélée vraiment efficace sur un ensemble de données réelles de 18902 compteurs électriques mesurant toutes les demi-heures la consommation électrique pendant deux semaines.

Mots clés : Estimateurs design-based, Composantes principales fonctionnelles, Consommation électrique, Estimateur d'Horvitz-Thompson.

∗. Compte-rendu de conférence écrit en collaboration avec Hervé Cardot et Alain Dessertaine, et présentée lors de la 19ème conférence de IASC-ERS (European Regional Section of the International Association for Statistical Computing), COMPSTAT'2010, et publié sous la référence suivante :
Cardot, H. Dessertaine, A. and Josserand, E. (2010). Semiparametric models with functional responses in a model assisted survey sampling setting. *Compstat 2010*, Eds Lechevallier, Y. and Saporta, G. Physica-Verlag, Springer, 411-420.

Semiparametric models with functional response in a survey sampling setting: model assisted estimation of electricity consumption curves

Hervé Cardot[1], Alain Dessertaine[2], and Etienne Josserand[1]

[1] Institut de Mathématiques de Bourgogne, UMR 5584 CNRS,
Université de Bourgogne, 9, Av. A. Savary - B.P. 47 870, 21078 Dijon, France
herve.cardot@u-bourgogne.fr, etienne.josserand@u-bourgogne.fr
[2] EDF, R&D, ICAME - SOAD,
1, Av. du Général de Gaulle, 92141 Clamart, France
alain.dessertaine@edf.fr

Abstract

This work adopts a survey sampling point of view to estimate the mean curve of large databases of functional data. When storage capacities are limited, selecting, with survey techniques a small fraction of the observations is an interesting alternative to signal compression techniques. We propose here to take account of real or multivariate auxiliary information available at a low cost for the whole population, with semiparametric model assisted approaches, in order to improve the accuracy of Horvitz-Thompson estimators of the mean curve. We first estimate the functional principal components with a design based point of view in order to reduce the dimension of the signals and then propose semiparametric models to get estimations of the curves that are not observed. This technique is shown to be really effective on a real dataset of 18902 electricity meters measuring every half an hour electricity consumption during two weeks.

Keywords: Design-based estimation, Functional Principal Components, Electricity consumption, Horvitz-Thompson estimator

5.1 Introduction

With the development of distributed sensors one can have access of potentially huge databases of signals evolving along fine time scales. Collecting in an exhaustive way such data would require very high investments both for transmission of the signals through networks as well as for storage. As noticed in Chiky and Hébrail (2009) survey sampling procedures on the sensors, which allow a trade off between limited storage capacities and accuracy of the data, can be relevant approaches compared to signal compression in order to get accurate approximations to simple estimates such as mean or total trajectories. Our study is motivated, in such a context of distributed data streams, by the estimation of the temporal evolution of electricity consumption curves. The French operator EDF has planned to install in a few years more than 30 millions electricity meters, in each firm and household, that will be able to send individual electricity consumptions at very fine time scales. Collecting, saving and analysing all this information which can be seen

as functional would be very expensive and survey sampling strategies are interesting to get accurate estimations at reasonable costs (Dessertaine, 2006). It is well known that consumption profiles strongly depend on covariates such as past consumptions, meteorological characteristics (temperature, nebulosity, *etc*) or geographical information (altitude, latitude and longitude). Taking this information into account at an individual level (*i.e* for each electricity meter) is not trivial.

We have a test population of $N = 18902$ electricity meters that have collected electricity consumptions every half an hour during a period of two weeks, so that we have $d = 336$ time points. We are interested in estimating the mean consumption curve during the second week and we suppose that we know the mean consumption, $\bar{Y}_k = \frac{1}{336}\sum_{j=1}^{336} Y_k(t_j)$, for each meter k of the population during the first week. This mean consumption will play the role of auxiliary information. Note that meteorological variables are not available in this preliminary study.

One way to achieve this consists in reducing first the high dimension of the data by performing a functional principal components analysis in a survey sampling framework with a design based approach (Cardot *et al.*, 2010). It is then possible to build models, parametric or nonparametric, on the principal component scores in order to incorporate the auxiliary variables effects and correct our estimator with model assisted approaches (Särndal *et al.*, 1992). Note that this strategy based on modeling the principal components instead of the original signal has already been proposed, with a frequentist point of view, by Chiou *et al.* (2003) with singel index models and Müller and Yao (2008) with additive models.

We present in section 2 the Horvitz-Thomposon estimator of the mean consumption profile as well as the functional principal components analysis. We develop, in section 3, model assisted approaches based on statistical modeling of the principal components scores and derive an approximated variance that can be useful to build global confidence bands. Finally, we illustrate, in section 4, the effectiveness of this methodology which allows to improve significantly more basic approaches on a population of 18902 electricity consumption curves measured every half an hour during one week.

5.2 Functional data in a finite population

Let us consider a finite population $U = \{1, \ldots, k, \ldots, N\}$ of size N, and suppose we can observe, for each element k of the population U, a deterministic curve $Y_k = (Y_k(t))_{t\in[0,1]}$ that is supposed to belong to $L^2[0,1]$, the space of square integrable functions defined on the closed interval $[0,1]$ equipped with its usual inner product $\langle\cdot,\cdot\rangle$ and norm denoted by $\|\cdot\|$. Let us define the mean population curve $\mu \in L^2[0,1]$ by

$$\mu(t) \quad = \quad \frac{1}{N}\sum_{k\in U} Y_k(t), \quad t \in [0,1]. \tag{5.1}$$

Consider now a sample s, *i.e.* a subset $s \subset U$, with known size n, chosen randomly according to a known probability distribution p defined on all the subsets of U. We suppose that all the individuals in the population can be selected, with probabilities that may be unequal, $\pi_k = \Pr(k \in s) > 0$ for all $k \in U$ and $\pi_{kl} = \Pr(k \,\&\, l \in s) > 0$ for all $k, l \in U$,

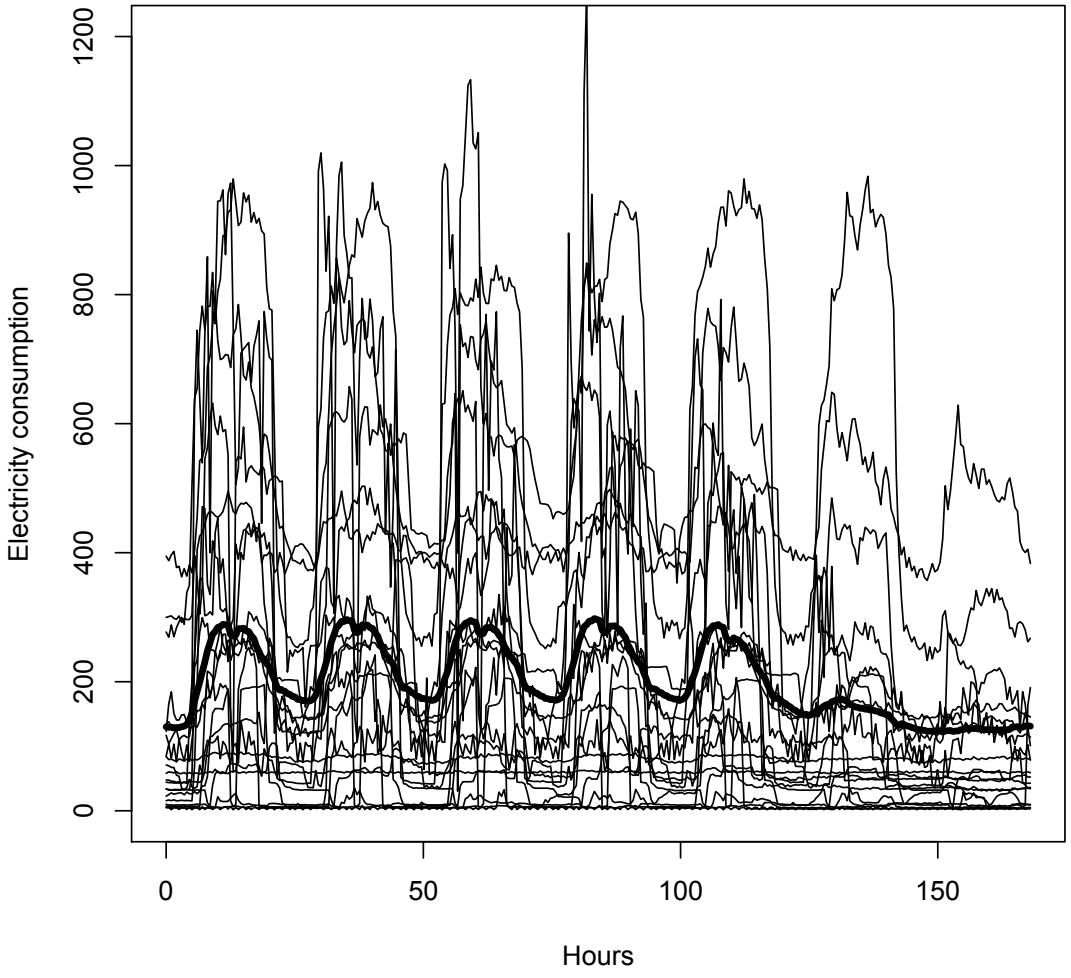

Figure 5.1: Mean curve and sample of individual electricity consumption curves.

$k \neq l$. The Horvitz-Thompson estimator of the mean curve, which is unbiased, is given by

$$\widehat{\mu}(t) \quad = \quad \frac{1}{N}\sum_{k\in s}\frac{Y_k(t)}{\pi_k} = \frac{1}{N}\sum_{k\in U}\frac{Y_k(t)}{\pi_k}\mathbf{1}_{k\in s}, \quad t\in[0,1]. \tag{5.2}$$

As in Cardot *et al.* (2010) we would like to describe now the individual variations around the mean function in a functional space whose dimension is as small as possible according to a quadratic criterion. Let us consider a set of q orthonormal functions of $L^2[0,1]$, $\phi_1,\ldots,\phi_q$, and minimize, according to $\phi_1,\ldots,\phi_q$, the remainder $R(q)$ of the projection of the Y_k's onto the space generated by these q functions

$$R(q) = \frac{1}{N}\sum_{k\in U}\|R_{qk}\|^2$$

with $R_{qk}(t) = Y_k(t) - \mu(t) - \sum_{j=1}^{q} \langle Y_k - \mu, \phi_j \rangle \phi_j(t),\ t \in [0,1]$. Introducing now the population covariance function $\gamma(s,t)$,

$$\gamma(s,t) = \frac{1}{N} \sum_{k \in U} \left(Y_k(t) - \mu(t)\right) \left(Y_k(s) - \mu(s)\right), \quad (s,t) \in [0,1] \times [0,1],$$

Cardot *et al.* (2010) have shown that $R(q)$ attains its minimum when $\phi_1, \ldots, \phi_q$ are the eigenfunctions of the covariance operator Γ associated to the largest eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q \geq 0$,

$$\Gamma \phi_j(t) = \int_0^1 \gamma(s,t) \phi_j(s) ds = \lambda_j \phi_j(t), \quad t \in [0,1], j \geq 1.$$

When observing individuals from a sample s, a simple estimator of the covariance function

$$\widehat{\gamma}(s,t) \quad = \quad \frac{1}{N} \sum_{k \in s} \frac{1}{\pi_k} \left(Y_k(t) - \widehat{\mu}(t)\right) \left(Y_k(s) - \widehat{\mu}(s)\right)\ (s,t) \in [0,1] \times [0,1], \qquad (5.3)$$

allows to derive directly estimators of the eigenvalues $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_q$ and the corresponding eigenfuctions $\widehat{\phi}_1, \ldots, \widehat{\phi}_q$.

Remark: *with real data, one only gets discretized trajectories of the Y_k at d points, $t_1, \ldots, t_d$, so that we observe $\mathbf{Y}_k = (Y_k(t_1), \ldots, Y_k(t_d)) \in \mathbb{R}^d$. When observations are not corrupted by noise, linear interpolation allows to get accurate approximations to the true trajectories,*

$$\widetilde{Y}_k(t) = Y_k(t_j) + \frac{Y_k(t_{j+1}) - Y_k(t_j)}{t_{j+1} - t_j}(t - t_j), \quad t \in [t_j, t_{j+1}]$$

and to build consistent estimates of the mean function provided the grid of time points is dense enough (Cardot and Josserand, 2009).

5.3 Semiparametric estimation with auxiliary information

Suppose now we have access to m auxiliary variables $X_1, \ldots, X_m$ that are supposed to be linked to the individual curves Y_k and we are able to observe these variables, at a low cost, for every individual k in the population. Taking this additional information into account would certainly be helpful to improve the accuracy of the basic estimator $\widehat{\mu}$. Going back to the decomposition of the individual trajectories Y_k on the eigenfunctions,

$$Y_k(t) = \mu(t) + \sum_{j=1}^{q} \langle Y_k - \mu, \phi_j \rangle \phi_j(t) + R_{qk}(t), \quad t \in [0,1],$$

and borrowing ideas from Chiou *et al.* (2003) and Müller and Yao (2008), an interesting approach consists in modeling the population principal components scores $\langle Y_k - \mu, \phi_j \rangle$ with respect to auxiliary variables at each level j of the decomposition on the eigenfunctions, $\langle Y_k - \mu, \phi_j \rangle \approx f_j(x_{k1}, \ldots, x_{km})$ where the regression function f_j can be parametric or not and $(x_{k1}, \ldots, x_{km})$ is the vector of observations of the m auxiliary variables for individual k.

It is possible to estimate the principal component scores

$$\widehat{C}_{kj} = \langle Y_k - \widehat{\mu}, \widehat{\phi}_j \rangle,$$

for $j = 1, \ldots, q$ and all $k \in s$. Then, a design based least squares estimator for the functions f_j

$$\widehat{f}_j \quad = \quad \arg\min_{g_j} \sum_{k \in s} \frac{1}{\pi_k} \left(\widehat{C}_{kj} - g_j(x_{k1}, \ldots, x_{km}) \right)^2, \tag{5.4}$$

is useful to construct the following model-assisted estimator $\widehat{\mu}_X$ of μ,

$$\widehat{\mu}_x(t) \quad = \quad \widehat{\mu}(t) - \frac{1}{N} \left(\sum_{k \in s} \frac{\widehat{Y}_k(t)}{\pi_k} - \sum_{k \in U} \widehat{Y}_k(t) \right) \tag{5.5}$$

where the predicted curves $\widehat{Y}_k$ are estimated for all the individuals of the population U thanks to the m auxiliary variables,

$$\widehat{Y}_k(t) = \widehat{\mu}(t) + \sum_{j=1}^{q} \widehat{f}_j(x_{k1}, \ldots, x_{km}) \ \widehat{v}_j(t), \ t \in [0, 1].$$

5.4 Estimation of electricity consumption curves

We consider now the population consisting in the $N = 18902$ electricity consumption curves measured during the second week very half an hour. We have $d = 336$ time points. Note that meteorological variables are not available in this preliminary study and our auxiliary information is the mean consumption, for each meter k, during the first week.

We first perform a simple random sampling without replacement (SRSWR) with fixed size of $n = 2000$ electricity meters during the second week in order to get $\widehat{\mu}$ and perform the functional principal components analysis (FPCA). The true mean consumption curve $\mu(t)$ during this period is drawn in Figure 5.1 whereas Figure 5.2 (a) present the result of the FPCA. The first principal component explains more than 80% of the total variance telling us that there is a strong temporal structure in these data. The associated estimated eigenfunction $\widehat{\phi}_1$ presents strong daily periodicity. Looking now at the relationship between the estimated first principal components and the auxiliary variable, we can notice that there is a strong linear relationship between these two variables and thus considering a linear regression model for estimating f_1 seems to be appropriate.

To evaluate the accuracy of estimator (5.5) we made 500 replications of the following scheme

- Draw a sample of size $n = 2000$ in population U with SRSWR and estimate $\widehat{\mu}, \widehat{\phi}_1$ and $\widehat{C}_{k1}$, for $k \in s$, during the second week.
- Estimate a linear relationship between X_k and $\widehat{C}_{k1}$, for $k \in s$ where $X_k = \frac{1}{336} \sum_{j=1}^{336} Y_k(t_j)$ is the mean consumption during the first week, and predict the principal component using the estimated relation $\widehat{C}_{k1} \approx \widehat{\beta}_0 + \widehat{\beta}_1 X_k$.
- Estimate $\widehat{\mu}_X$ taking the auxiliary information into account with equation (5.5).

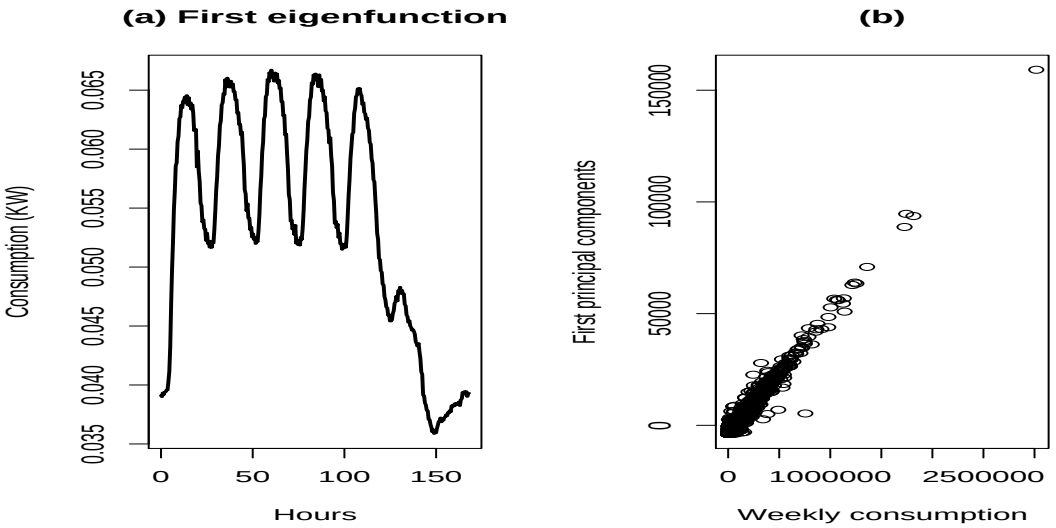

Figure 5.2: Mean curve and sample of individual electricity consumption curves.

	SAS	OPTIM	MA1
Mean	4.245	1.687	1.866
Median	3.348	1.613	1.813
First quartile	2.213	1.343	1.499
Third quartile	5.525	1.944	2.097

Table 5.1: Comparison of mean absolute deviation from the true mean curve for SRSWR, optimal allocation for stratification (OPTIM) and model assisted (MA1) estimation procedures.

The following loss criterion $\int |\mu(t) - \widehat{\mu}(t)|dt$ has been considered to evaluate the accuracy of the estimators $\widehat{\mu}$ and $\widehat{\mu}_X$. We also compare the estimation error with an optimal stratification sampling scheme in which strata are built on the curves of the population observed during the first week. As in Cardot and Josserand (2009), the population is partitioned into $K = 7$ strata thanks to a k-means algorithm. It is then possible to determine the optimal allocation weights, according to a mean variance criterion, in each stratum for the stratified sampling procedure during the second week.

The estimation errors are presented in Table 5.1 for the three estimators. We first remark that considering optimal stratification (OPTIM) or model assisted estimators (MA1) lead to a significant improvement compared to the basic SRSWR approach. Secondly, the performances of the stratification and the model assisted approaches are very similar in terms of accuracy but they do not need the same amount of information. The optimal stratification approach necessitates to know the cluster of each individual of the population and the covariance function within each cluster whereas the model assisted estimator only needs the past mean consumption for each element of the population.

Looking now at the empirical variance, at each instant, of these estimators, we see in Figure (5.3) that the simple SRSWR has much larger variances, in which we recognize the

first eigenfunction of the covariance operator, than the more sophisticated OPTIM and MA1. Among these two estimators the model assisted estimator has a smaller pointwise variance, indicating that it is certainly more reliable.

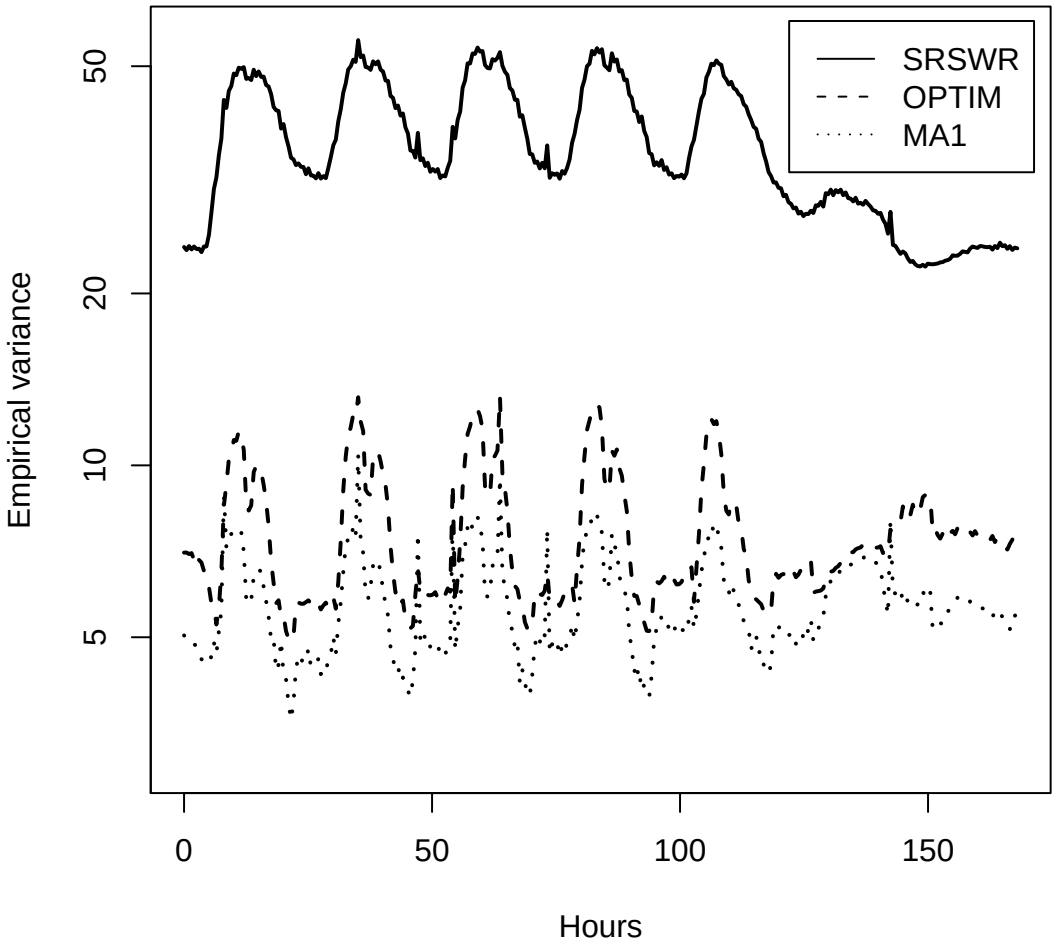

Figure 5.3: Comparison of the empirical pointwise variance for SRSWR, optimal allocation for stratification (OPTIM) and model assisted (MA1) estimation procedures.

Acknowledgment. Etienne Josserand thanks the Conseil Régional de Bourgogne, France, for its financial support (FABER PhD grant).

References

CARDOT, H., CHAOUCH, M., GOGA, C. and C. LABRUÈRE (2010). Properties of Design-Based Functional Principal Components Analysis. *J. Statist. Planning and*

Inference., **140**, 75-91.

CARDOT, H., JOSSERAND, E. (2009). Horvitz-Thompson Estimators for Functional Data: Asymptotic Confidence Bands and Optimal Allocation for Stratified Sampling. `http://arxiv.org/abs/0912.3891`.

CHIKY, R., HEBRAIL, G. (2009). Spatio-temporal sampling of distributed data streams. *J. of Computing Science and Engineering*, to appear.

CHIOU, J-M., MÜLLER, H.G. and WANG, J.L. (2003). Functional quasi-likelihood regression models with smooth random effects. *J.Roy. Statist. Soc., Ser. B*, **65**, 405-423.

DESSERTAINE, A. (2006). Sampling and Data-Stream : some ideas to built balanced sampling using auxiliary hilbertian informations. *56th ISI Conference*, Lisboa, Portugal, 22-29 August 2007.

MÜLLER, H-G., YAO, F. (2008). Functional Additive Model. *J. Am. Statist. Ass.* **103**, 1534-1544.

SÄRNDAL, C.E., SWENSSON, B. and J. WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

SKINNER, C.J, HOLMES, D.J, SMITH, T.M.F (1986). The Effect of Sample Design on Principal Components Analysis. *J. Am. Statist. Ass.* **81**, 789-798.

Chapitre 6

Estimation and confidence bands for the mean electricity consumption curve : a comparison of unequal probability sampling designs and model assisted approaches *

Résumé : Quand les capacités de stockage sont limitées ou que les coûts de transmission sont élevés, sélectionner grâce à des techniques de sondage une petite fraction des observations est une alternative intéressante aux techniques de compression du signal. Notre étude est motivée, dans ce contexte, par l'estimation de l'évolution temporelle de la moyenne des courbes de consommation électrique. Nous comparons dans ce travail différentes manières de prendre en compte l'information auxiliaire. La première consiste à utiliser des plans simples, comme le sondage aléatoire simple sans remise, et les estimateurs par modèle assisté obtenu en réduisant d'abord la dimension de la variable d'intérêt fonctionnelle. Une seconde stratégie consiste à considérer des plans à probabilités inégales comme le sondage stratifié ou πps qui peut prendre en compte des informations supplémentaires directement dans les poids d'échantillonnage. La deuxième question abordée dans ce travail est la construction de bandes de confiance fiables. Lorsque des estimateurs consistants de la fonction de covariance des estimateurs sont faciles à construire et que l'estimateur de la moyenne satisfait un Théorème Central Limite fonctionnel, une technique rapide basée sur des simulations de processus gaussiens, afin d'approcher la distribution de leurs suprema, peut être mis en place. Cette nouvelle approche est comparée à des techniques de bootstrap qui sont également des candidats naturels pour la construction de bandes de confiance et qui peuvent être adaptées aux paramètres de la population finie.

Mots clés : Bootstrap, Estimateurs design-based, Composantes principales fonctionnelles, Estimateur d'Horvitz-Thompson, Supremum de processus Gaussien, Échantillonnage.

∗. Travaux effectués avec la collaboration de Hervé Cardot, Alain Dessertaine et de Pauline Lardin, et dont les résultats seront présentés lors du 58ème congrès de l'ISI (International Statistical Institute).

Estimation and confidence bands for the mean electricity consumption curve: a comparison of unequal probability sampling designs and model assisted approaches

Hervé Cardot[1], Alain Dessertaine[2], Etienne Josserand[1] and Pauline Lardin[1,2]

[1] Institut de Mathématiques de Bourgogne, UMR 5584 CNRS,
Université de Bourgogne, 9, Av. A. Savary - B.P. 47 870, 21078 Dijon, France
herve.cardot@u-bourgogne.fr, etienne.josserand@u-bourgogne.fr
[2] EDF, R&D, ICAME - SOAD,
1, Av. du Général de Gaulle, 92141 Clamart, France
alain.dessertaine@edf.fr, pauline.lardin@edf.fr

Abstract

When 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 context, by the estimation of the temporal evolution of mean electricity consumption curves. 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 performed by first reducing the dimension of the functional variable of interest. 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. The second question addressed in this work is how to build 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 new 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.

Keywords: Bootstrap, Design-based estimation, Functional Principal Components, Horvitz-Thompson estimator, Supremum of Gaussian processes, Survey sampling.

6.1 Introduction

We consider a survey sampling point of view in order to estimate the mean curve of large databases of functional data. When 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 context, by the estimation of the temporal evolution of mean electricity consumption

curves. The French operator EDF has planned to install in a few years more than 30 millions electricity meters, in each firm and household, that will be able to send individual electricity consumptions at very fine time scales. Collecting, saving and analyzing all this information which can be seen as functional would be very expensive and survey sampling strategies are interesting to get accurate estimations at reasonable costs (Dessertaine, 2008). It is also well known that consumption profiles may depend on covariates such as past aggregated consumptions, meteorological characteristics (temperature, etc) or geographical information (altitude, latitude and longitude).

We compare in this work different ways of taking this information into account. A first one consists in using simple sampling designs, such as simple random sampling without replacement, and model assisted estimators (Särndal et al. 1992). 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.

The second question addressed in this work is how to build 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 (Cardot and Josserand, 2011), a fast technique, inspired from Degras (2011), based on simulations of Gaussian processes in order to approximate the distribution of their suprema can be employed. This new 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 (Booth et al. 1994, Chauvet, 2007).

6.2 Functional data in a finite population

Let us consider a finite population $U = \{1, \ldots, k, \ldots, N\}$ of size N, and suppose we can observe, for each element k of the population U, a deterministic curve $Y_k = (Y_k(t))_{t\in[0,T]}$ that is supposed to belong to $C[0,T]$, the space of continuous functions defined on the closed interval $[0,T]$. Let us define the mean population curve $\mu \in C[0,T]$ by

$$\mu(t) \quad = \quad \frac{1}{N}\sum_{k\in U} Y_k(t), \quad t \in [0,T]. \tag{6.1}$$

Consider now a sample s, *i.e.* a subset $s \subset U$, with known size n, chosen randomly according to a known probability distribution p defined on all the subsets of U. We suppose that all the individuals in the population can be selected, with probabilities that may be unequal, $\pi_k = \Pr(k \in s) > 0$ for all $k \in U$ and $\pi_{kl} = \Pr(k \ \& \ l \in s) > 0$ for all $k, l \in U$, $k \neq l$.

The Horvitz-Thompson estimator of the mean curve (Cardot et al. 2010), which is unbiased, is given by

$$\widehat{\mu}(t) \quad = \quad \frac{1}{N}\sum_{k\in s} \frac{Y_k(t)}{\pi_k} = \frac{1}{N}\sum_{k\in U} \frac{Y_k(t)}{\pi_k}\mathbb{1}_{k\in s}, \quad t \in [0,T]. \tag{6.2}$$

In this context, we can define $\widehat{\mu}_{\text{srswor}}$, the simple random sampling without replacement mean estimator, by

$$\widehat{\mu}_{\text{srswor}}(t) \quad = \quad \frac{1}{n}\sum_{k\in s} Y_k(t), \quad t \in [0,T]. \tag{6.3}$$

6.3 Estimators using auxiliary information

We consider now the particular case of stratified sampling with simple random sampling without replacement in all strata, assuming the population U is divided into a fixed number H of strata. This means that there is a partitioning of U into H subpopulations denoted by U_h, $(h = 1, \ldots, H)$. We can define the mean curve μ_h within each stratum h as $\mu_h(t) = N_h^{-1} \sum_{k \in U_h} Y_k(t),\ t \in [0, T]$, where N_h is the number of units in stratum h. The first and second order inclusion probabilities are explicitly known, and the mean curve estimator of $\mu_N(t)$ is

$$\widehat{\mu}_{\text{strat}}(t) \quad = \quad \frac{1}{N} \sum_{h=1}^{H} n_h^{-1} N_h \sum_{k \in s_h} Y_k(t),\ t \in [0, T], \tag{6.4}$$

where s_h is a sample of size n_h, with $n_h \leq N_h$, obtained by simple random sampling without replacement in stratum U_h.

Auxiliary information can be taken into account to build strata in order to improve the accuracy of the mean estimator. The sample size n_h in stratum h is determined by a Neyman-like allocation, as suggested in Cardot and Josserand (2011), in order to get a Horvitz-Thompson estimator of the mean trajectory whose variance is as small as possible.

Another interesting sampling design is the πps which can use directly auxiliary information. Indeed, we defined the first inclusion probability by

$$\pi_k = n \frac{x_k}{\sum_{k \in U} x_k}, \tag{6.5}$$

where x_k is a real auxiliary variable for these k. For some units, π_k can be higher than one. To carry out this problem, we select automatically these units. Then, we compute again the first inclusion probabilities without the units already selected. We repeat this algorithm until all π_k are lower or equal to one. Using (6.2), we then obtain the πps mean estimator $\widehat{\mu}_{\pi\text{ps}}$.

Instead of using the auxiliary information into the sampling design, we can adjust a linear model and build a model assisted estimator $\widehat{\mu}_{\text{ma}}$. More precisely, we can write for all units k and $t \in [0, T]$

$$Y_k(t) = \beta_0(t) + \beta_1(t) x_k + \epsilon_{kt} \tag{6.6}$$

where $\beta_0(t)$ and $\beta_1(t)$ are regression coefficients (see Faraway, 1997). Survey sampling weights are taken into account to compute the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ of β_0 and β_1 (see Särndal et al. 1992). Finally, we get the mean estimator, for $t \in [0, T]$,

$$\widehat{\mu}_{\text{ma}}(t) = \frac{1}{N} \sum_{k \in s} \frac{Y_k(t)}{\pi_k} - \frac{1}{N} \left(\sum_{k \in s} \frac{\widehat{Y}_k(t)}{\pi_k} - \sum_{k \in U} \widehat{Y}_k(t) \right) \tag{6.7}$$

where $\widehat{Y}_k(t) = \widehat{\beta}_0(t) + \widehat{\beta}_1(t) x_k,\ t \in [0, T]$.

6.4 Confidence bands

In this section, we want to build confidence bands for μ of the form

$$\left\{ \left[\widehat{\mu}(t) \pm c\, \widehat{\sigma}(t) \right],\ t \in [0, T] \right\}, \tag{6.8}$$

where c is a suitable number and $\widehat{\sigma}(t)$ is an estimator of $\gamma(t,t)^{1/2}$, and where $\gamma(s,t) = \mathrm{Var}(\widehat{\mu}(s), \widehat{\mu}(t))$ is the covariance function of $\widehat{\mu}$. More precisely, given a confidence level $1-\alpha \in]0,1[$, we seek $c = c_\alpha$ that satisfies approximately

$$\mathbb{P}\left(\mu \in \{[\widehat{\mu}(t) \pm c_\alpha\, \widehat{\sigma}(t)]\,,\ \forall t \in [0,T]\}\right) = 1-\alpha. \tag{6.9}$$

6.4.1 Suprema of Gaussian processes

We consider the process $Z(t) = \big(\widehat{\mu}(t) - \mu(t)\big)/\widehat{\sigma}(t)$ which converges to a Gaussian process in the space of continuous functions $\mathcal{C}([0,T])$, under some technical assumptions (Cardot and Josserand, 2011). We can determine c_α such that

$$\mathbb{P}\left(|Z(t)| \leq c_\alpha,\ \forall t \in [0,T]\right) = 1-\alpha, \tag{6.10}$$

where Z is a Gaussian process with mean zero and correlation function ρ, and where $\rho(s,t) = \widehat{\gamma}(s,t)/\big(\widehat{\gamma}(s,s)\ \widehat{\gamma}(t,t)\big)^{1/2}$. Note that the calculus of c_α with a Gaussian process is only possible when one can build an estimator $\widehat{\gamma}$ of the covariance function γ.

6.4.2 Bootstrap bands

Another way consists in estimating the covariance function by bootstrap (Booth et al. 1994, Chauvet, 2007). Using the sample s, we can generate a fictive population $U^\star$ and by simulation we obtain an approximation of $\widehat{\sigma}$. The following algorithm permits to build confidence bands:

1. Draw a sample s, with known size n, chosen randomly according to a known probability distribution p, and to compute $\widehat{\mu}$.
2. Duplicate each units $k \in s$ $1/\pi_k$ times to build a fictive population $U^\star$.
3. Draw in $U^\star$ M samples $s_j^\star$ with size n according to p, and to generate $\widehat{\mu}_j^\star(t)$, $j = 1, \ldots, M$.
4. The function $\widehat{\sigma}(t)$ is estimated by the empirical standard deviation of $\widehat{\mu}_j^\star(t)$, $j = 1, \ldots, M$.
5. Let $E_{c_\alpha} = \{j | \forall t \quad \widehat{\mu}(t) \in [\widehat{\mu}_j^\star(t) - c_\alpha \widehat{\sigma}(t); \widehat{\mu}_j^\star(t) + c_\alpha \widehat{\sigma}(t)]\}$. The coefficient c_α is chosen such that $\#(E_{c_\alpha}) = (1-\alpha)M$.

The second step of this algorithm may causes some problems because $1/\pi_k$ is not necessarily an integer. This is detailed in the next section and was already discussed by Booth *et al* (1994) and Chauvet (2007).

6.5 Study of mean electricity consumption curve

We consider now a population consisting in the $N = 15069$ electricity consumption curves measured during one week every half an hour. We have $d = 336$ time points. Note that our auxiliary information is the mean consumption, for each meter k, during previous week. We compare previous estimators with fixed size $n = 1500$. For each estimator we compute the confidence bands with the Gaussian bands and the bootstrap bands procedures. A draw back of Gaussian bands is that they require a covariance function estimator $\widehat{\gamma}$ whereas bootstrap methods just needs some adjustment to build a fictive population.

– $\widehat{\mu}_{\text{srswor}}$: For the simple random sampling without replacement estimator, we have an unbiased covariance function estimator

$$\widehat{\gamma}_{\text{srswor}}(s,t) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(\frac{1}{n-1}\sum_{k,l\in s} Y_k(s)Y_l(t) - \frac{n}{n-1}\,\widehat{\mu}_{\text{srswor}}(s)\,\widehat{\mu}_{\text{srswor}}(t)\right),\ s,t\in[0,T]. \tag{6.11}$$

To build the fictive population in the bootstrap step 2, we can remark that $1/\pi_k = N/n$ is not an integer. So, we duplicate $k \in s$ $[N/n]$ times, where [.] is the entire part function. We complete the duplication step with a simple random sampling without replacement in s with a fixed size $N - n[N/n]$, in order to obtain a fictive population $U^\star$ which the size is equal to N.

– $\widehat{\mu}_{\text{strat}}$: The population is partitioned into $H = 10$ strata thanks to a k-means algorithm on our auxiliary variable, the mean consumption during the first week. The covariance function is estimated by

$$\widehat{\gamma}_{\text{strat}}(s,t) = \frac{1}{N^2}\sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h}\,\widehat{\gamma}_h(s,t)\ \ s,t\in[0,T], \tag{6.12}$$

where $\widehat{\gamma}_h$ is covariance function estimator into stratum h.
The fictive population $U^\star$ is obtained by the same method used for $\widehat{\mu}_{\text{srswor}}$ in each stratum h.

– $\widehat{\mu}_{\pi\text{ps}}$: With the πps, it is difficult to obtain a formula for second inclusion probabilities because they depend on how the sample is drawn and there is no standard method. When the sample size is fixed and the sampling design p is close to the maximal entropy, we can use the Hájek approximation (Hájek, 1964) which can be adapted to get the following estimation of the covariance function

$$\widehat{\gamma}_{\pi\text{ps}}(s,t) = \frac{1}{N^2}\sum_{k\in s}(1-\pi_k)\left(\frac{Y_k(t)}{\pi_k} - \widehat{R}(t)\right)\left(\frac{Y_k(s)}{\pi_k} - \widehat{R}(s)\right)\ \ s,t\in[0,T], \tag{6.13}$$

where $\widehat{R}(t) = \sum_{k\in s}\frac{Y_k(t)}{\pi_k}(1-\pi_k)/\sum_{k\in s}(1-\pi_k)$.
For the bootstrap, each $k \in s$ is duplicated $[1/\pi_k]$ times. As suggested in Chauvet (2007), to complete the population $U^\star$, we realize a πps sampling with an inclusion probability $\alpha_k = 1/\pi_k - [1/\pi_k]$. To keep a fixed sample size during the bootstrap step 3, we use the sampling design $p^\star$ defined for all $k \in U^\star$ by

$$\pi_k^\star = n\frac{x_k}{\sum_{k\in U^\star} x_k}. \tag{6.14}$$

– $\widehat{\mu}_{\text{ma}}$: The covariance function of the model assisted estimator is complicated to explicit because it depends on the sampling design and the model. By analogy with Breidt and Opsomer (2000), we have an asymptotic covariance estimator

$$\widehat{\gamma}_{\text{ma}}(s,t) = \frac{1}{N^2}\sum_{k,l\in s}\left(Y_k(s) - \widehat{Y}_k(s)\right)\left(Y_l(t) - \widehat{Y}_l(t)\right)\frac{\pi_{kl} - \pi_k\pi_l}{\pi_k\pi_l\pi_{kl}}\ \ s,t\in[0,T], \tag{6.15}$$

where $\pi_k = \frac{n}{N}$ et $\pi_{kl} = \frac{n(n-1)}{N(N-1)}$ for $k,l \in s$ and $k \neq l$. To build the bootstrap bands, we adapt an algorithm of Helmers and Wegkamp (1998) to our case. For

each sample s, we have $\widehat{\epsilon}_{kt} = \widehat{Y}_k(t) - \widehat{\beta}_0(t) - \widehat{\beta}_1(t)x_k$ for all $k \in s$. We then draw n *iid* realizations $Z_1, \ldots, Z_n$ of a centered gaussian variable with unit variance and consider

$$Y_k^\star(t) = \widehat{\beta}_0(t) + \widehat{\beta}_1(t)x_k + Z_k\widehat{\epsilon}_{kt} \;\; t \in [0,T]. \tag{6.16}$$

By a simple random sampling without replacement in the fictive population $U^\star$, we obtain the mean estimator for $s^\star$

$$\widehat{\mu}_{\text{ma}}^\star(t) = \frac{1}{N}\sum_{k\in U}\tilde{Y}_k(t) - \frac{1}{N}\sum_{k\in s^\star}\frac{\tilde{Y}_k(t) - Y_k(t)}{\pi_k} \;\; t \in [0,T]. \tag{6.17}$$

where $\tilde{Y}_k(t) = \widehat{\beta}_0^\star(t) + \widehat{\beta}_1^\star(t)x_k$ for all $k \in U$, $\widehat{\beta}_0^\star$ and $\widehat{\beta}_1^\star$ are model parameters computing on $s^\star$, and $\pi_k = n/N$ for all $k \in s^\star$.

Afer computing the simulations, we present the covering rate of the confidence bands in Table 6.1 and the mean areas of the confidence bands in Table 6.2.

Methods	Number of simulations	Number of processes	Bootstrap		Gaussian process	
			α=0.05	α=0.01	α=0.05	α=0.01
SRSWOR	10000	10000	95.73	98.92	94.02	98.32
MA	500	5000	94.20	98.80	94.40	98.20
STRAT	2000	5000	93.80	98.25	94.00	98.30
πps	1169	1000	94.70	98.55	94.10	98.55

Table 6.1: Empirical coverage rates in percentage.

Methods	Number of simulations	Number of processes	Bootstrap		Gaussian process	
			α=0.05	α=0.01	α=0.05	α=0.01
SRSWOR	10000	10000	37.83	45.72	35.89	43.08
MA	500	5000	20.16	23.06	19.84	22.53
STRAT	2000	5000	16.64	18.92	16.61	18.88
πps	1169	1000	17.86	20.32	17.63	19.95

Table 6.2: Mean areas of the confidence bands.

As results, we note that both methods employed to build the confidence bands give almost the same empirical coverage and are close to the nominal levels of confidence. Moreover, confidence bands areas are very close too. The Gaussian processes simulation bands are much faster to compute but require a reliable estimator of the covariance function. On the other hand, the bootstrap bands can be long to generate but need nothing more than the basic estimator. The best estimators are obtained with the stratified and πps designs.

Acknowledgment. Etienne Josserand thanks the *Conseil Régional de Bourgogne, France* for its financial support (FABER PhD grant).

References

Berger, Y. G. (1998). Rate of convergence for asymptotic variance of the Horvitz–Thompson estimator. *J. Statist. Planning and Inference*, **74**, 149-168.

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

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, **28**, 1026-1053.

Cardot, H., Chaouch, M., Goga, C. end C. Labruère (2010). Properties of Design-Based Functional Principal Components Analysis. *J. Statist. Planning and Inference*, **140**, 75-91.

Cardot, H., Josserand, E. (2011). Horvitz-Thompson Estimators for Functional Data: Asymptotic Confidence Bands and Optimal Allocation for Stratified Sampling. *Biometrika*, **98**, 107-118.

Chauvet, G. (2007). Méthodes de Bootstrap en population finie. *PhD Thesis*, Université Rennes II, France.

Degras, D. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica*, to appear.

Dessertaine, A. (2008). Estimation de courbes de consommation électrique à partir de mesures non synchrones. In *Méthodes de sondage*, Eds. Guibert, P., Haziza, D., Ruiz-Gazen, A. and Tillé, Y. Dunod, Paris, 353-357.

Faraway, J. (1997). Regression Analysis for a Functional Responses. *Technometrics*, **39**, 254-261.

Helmers, R., Wegkamp, M. (1998). Wild Bootstrapping in Finite Population with Auxiliary Information. *Scandinavian journal of statistics*, **25**, 383-399.

Särndal, C.E., Swensson, B. and J. Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

Chapitre 7

Conclusion et Perspectives

Dans cette thèse, nous nous sommes intéressés à l'estimation dans le cas de données fonctionnelles de quantités simples telles que la courbe moyenne. L'approche originale de combiner des techniques de sondage dans le cadre fonctionnel s'avère plutôt efficace dans la pratique. Elle permet de prendre en compte la façon dont sont collectées les données et d'attribuer naturellement un poids à chaque unité statistique. Nous avons proposé un estimateur d'Horvitz-Thompson de la courbe moyenne, et grâce à des hypothèses asymptotiques sur le plan de sondage nous avons établi un Théorème Central Limite fonctionnel dans le cadre des fonctions continues afin d'obtenir des bandes de confiance asymptotiques. Pour un plan d'échantillonnage à taille fixe, nous avons vu que le sondage stratifié peut grandement améliorer l'estimation comparativement au sondage aléatoire simple. De plus, nous avons étendu la règle d'allocation optimale de Neyman dans le contexte fonctionnel. La prise en compte d'information auxiliaire a été développée grâce à des estimateurs de type *model assisted*, mais aussi en utilisant directement cette information dans les poids d'échantillonnage avec le sondage à probabilités inégales proportionnelles à la taille. Par la suite, ce travail a été généralisé à un problème cher aux données fonctionnelles, celui des courbes bruitées. La simple interpolation linéaire réalisée sur les données à temps discret n'est plus vraiment justifiée dans ce contexte, et fut remplacée par un lissage par polynômes locaux. Les propriétés de consistance de nos estimateurs furent établies, ainsi que la normalité asymptotiques des estimateurs de la courbe moyenne. Deux méthodes de constructions des bandes de confiance ont été proposées. La première utilise la normalité asymptotique de nos estimateurs en simulant un processus Gaussien conditionnellement à sa fonction de covariance afin d'en estimer la loi du sup. La seconde utilise des techniques de bootstrap en population finie qui ne nécessite pas l'estimation de la fonction de covariance. La construction des bandes de confiance *via* un processus Gaussien demande une estimation de la covariance de l'estimateur. Des formules ont été établies dans le cas du sondage aléatoire simple sans remise et du sondage stratifié, et des approximations ont été obtenues pour les estimateurs de type *model assisted* et à probabilités inégales proportionnelles à la taille. Ces approximations demanderaient à être justifiées théoriquement en adaptant les résultats de Breidt & Opsomer (2000) et en démontrant la validité de la formule d'Hàjek utilisée dans le chapitre 6.

Le développement de ces méthodes a été motivé par une question industrielle, à savoir l'estimation de courbes de consommation électrique fournies par E.D.F. (Électricité De France), et nous a permis de garder à l'esprit une validité pratique de nos résultats. Néan-

moins, ces outils restent généraux et peuvent être mis en place dans d'autres domaines, comme par exemple dans des problématiques de réseaux où la quantité d'information en circulation est limitée. Une poursuite intéressante de ces travaux serait de faire du sondage fonctionnel sur données fonctionnelles. C'est-à-dire d'avoir des poids de sondage qui dépendent du temps et donc d'avoir la possibilité de changer d'échantillon à chaque instant ou chaque période de temps. La construction de tels estimateurs reste néanmoins complexe puisqu'il faut établir comment évoluent les poids de sondage au cours du temps. Les propriétés de base, comme le calcul de la variance ou la convergence asymptotique de tels estimateurs, ne sont pas évidentes et restent un problème ouvert.

Annexe

Annexe A

Outils Probabilistes

Les résultats asymptotiques présentées dans cette thèse reposent essentiellement sur des critères de tension des trajectoires ainsi que sur des propriétés d'inégalités maximales. Ces techniques sont utilisées pour montrer la consistance de nos estimateurs mais également la normalité des estimateurs de la courbe moyenne *via* un théorème central limite fonctionnel.

A.1 Tension

En reprenant les notations de Billingsley (1968), on note $C = C[0,T]$ l'ensemble des fonctions continues sur l'intervalle $[0,T]$ muni de la norme sup et $\mathcal{C}$ l'ensemble des boréliens de C.

Définition A.1.1. *Une mesure de probabilité P sur $(C,\mathcal{C})$ est dite tendue si pour tout réel ϵ strictement positif il existe un compact K tel que*

$$P(K) > 1 - \epsilon. \tag{A.1}$$

Une famille Π de mesures de probabilité sur l'espace C est dite tendue si pour tout réel ϵ strictement positif il existe un compact K tel que

$$P(K) > 1 - \epsilon, \quad \textit{pour tout } P \in \Pi. \tag{A.2}$$

On peut à présent donner le théorème principal qui justifiera le caractère gaussien asymptotique de nos estimateurs dans les théorèmes 2.3.3 et 4.3.1.

Théorème A.1.1 (Theorem 8.1 p 54, Bilingsley (1968))**.** *Soit P_n, P des mesures de probabilités sur $(C,\mathcal{C})$. Si les distributions finies dimensionnelles de P_n convergent faiblement vers celles de P, et si $\{P_n\}$ est tendue, alors P_n converge faiblement vers P dans C.*

Dans notre cas, on appliquera ce théorème à la suite $X_n(t) = \sqrt{n}(\widehat{\mu}_N(t) - \mu_N(t))$ avec $t \in [0,T]$. La définition de tension étant adaptée comme suit.

Définition A.1.2. *Si X_n sont des éléments aléatoires de C, on dit que $\{X_n\}$ est tendue quand $\{P_n\}$ est tendue, où P_n représente la distribution de X_n.*

La définition (A.1.1) n'est pas facile à manipuler. Dans la pratique, on prouvera le caractère de tension avec le théorème suivant.

Théorème A.1.2 (Theorem 12.3 p 95, Bilingsley (1968)). *La suite* $\{X_n\}$ *est tendue si elle satisfait ces deux conditions :*

i *La suite* $\{X_n(0)\}$ *est tendue.*

ii *Il existe des constantes* $\gamma \geq 0$ *et* $\alpha \geq 0$*, et une fonction* F *continue et non décroissante sur* $[0, T]$ *telle que*

$$\mathbb{P}\{|X_n(t_2) - X_n(t_1)| \geq \lambda\} \quad \leq \quad \frac{1}{\lambda^\gamma}|F(t_2) - F(t_1)|^\alpha \tag{A.3}$$

pour tous t_1*,* t_2 *et* n *et tout* λ *positif.*

La condition sur les moments

$$E\{|X_n(t_2) - X_n(t_1)|^\gamma\} \leq |F(t_2) - F(t_1)|^\alpha \tag{A.4}$$

implique (A.3).

C'est notamment l'équation (A.4) que l'on s'efforcera de montrer. L'intérêt de cette condition dans notre cadre sondage est qu'elle n'impose aucune condition d'indépendance et porte uniquement sur des propriétés de moments des accroissements. Pour obtenir ce genre d'inégalités, on utilisera des outils basés sur les inégalités maximales.

A.2 Inégalité maximale

Le livre de van der Vaart & Wellner (2000) consacre une section entière (section 2.2 p95) sur les inégalités maximales qui sont habituellement utilisées pour prouver l'équicontinuité asymptotique de processus empiriques.

Soit ψ une fonction non décroissante convexe avec $\psi(0) = 0$ et X une variable aléatoire, la norme d'Orlicz $\|X\|_\psi$ est définie par

$$\|X\|_\psi = \inf\left\{c > 0 : E\psi\left(\frac{|X|}{c}\right) \leq 1\right\}. \tag{A.5}$$

L'exemple le plus connu de norme d'Orlicz est celui correspondant aux fonctions de la forme $x \mapsto x^p$ avec $p \geq 1$. On retrouve simplement la norme L^p

$$\|X\|_\psi = (E|X|^p)^{1/p}. \tag{A.6}$$

Un résultat classique sur les inégalités maximales est que pour une certaine constante K dépendant de ψ, et pour des variables aléatoires quelconque $X_1, \ldots, X_m$, on a

$$\left\|\max_{1\leq i\leq m} X_i\right\|_\psi \leq K\psi^{-1}(m) \max_{1\leq i\leq m} \|X_i\|_\psi. \tag{A.7}$$

Le but est étendre ce genre de résultat pour contrôler une borne supérieure. L'idée consiste à relier les ensembles de maxima à un ensemble de maxima d'incréments.

Considérons, pour un processus stochastique $\{X_t, t \in [0, T]\}$, la semi-métrique

$$d_O(s, t) = \|X_s - X_t\|_\psi. \tag{A.8}$$

Définition A.2.1 (Covering numbers)**.** *Soit* $([0,T],d)$ *un espace semi-métrique. Alors le* covering number $N(\epsilon,d)$ *est le nombre minimal de boules de rayon* ϵ *nécessaire pour recouvrir* $[0,T]$*. On appelle une collection de points* ϵ*-separés si la distance entre chaque paire est strictement plus grande que* ϵ*. Le* packing number $D(\epsilon,d)$ *est le nombre maximum de points* ϵ*-separés dans* $[0,T]$

Les *covering numbers* est un bon moyen pour mesurer la finesse de la topologie engendrée par la semi-métrique d. Dans notre cas, pour les résultats qui suivent, les *covering number* et *packing number* peuvent être intervertis en remarquant que

$$N(\epsilon,d) \leq D(\epsilon,d) \leq N(\frac{1}{2}\epsilon,d). \tag{A.9}$$

On peut à présent énoncer le résultat principal qui nous permettra de prouver la tension.

Théorème A.2.1 (Theorem 2.2.4 p 98, van der Vaart & Wellner (2000))**.** *Soit* ψ *une fonction convexe, non décroissante, non nulle avec* $\psi(0)=0$ *et* $\limsup_{x,y\to\infty}\psi(x)\psi(y)/\psi(cxy) < \infty$*, pour une certaine constante* c*. Soit* $\{X_t, t\in[0,T]\}$ *un processus stochastique séparable avec*

$$\|X_s - X_t\|_\psi \leq C\ d(s,t),\ \textit{pour tout } s,t\in[0,T] \tag{A.10}$$

pour une certaine semi-métrique d *et une constante* C*. Alors, pour tout* $\eta,\delta>0$*,*

$$\left\| \sup_{d(s,t)\leq\delta} |X_s - X_t| \right\|_\psi \leq K\left[\int_0^\eta \psi^{-1}\left(D(\epsilon,d)\right)\ d\epsilon + \delta\psi^{-1}\left(D^2(\eta,d)\right)\right], \tag{A.11}$$

pour une constante K *dépendant seulement de* ψ *et de* C*.*

Corollaire A.2.1 (Corollary 2.2.5 p 98, van der Vaart & Wellner (2000))**.** *La constante* K *peut être choisie telle que*

$$\left\| \sup_{s,t} |X_s - X_t| \right\|_\psi \leq K\int_0^T \psi^{-1}\left(D(\epsilon,d)\right)\ d\epsilon. \tag{A.12}$$

En effet, on utilisera par la suite la fonction $\psi(x)=x^2$ ainsi qu'une semi-métrique de la forme

$$d(s,t) = E\left\{|X_n(s) - X_n(t)|^2\right\} \tag{A.13}$$

pour prouver la tension de nos trajectoires *via* (A.4). Le théorème A.2.1 joue un rôle important pour montrer la consistance des différents estimateurs présentés. Il permet de contrôler les différences du processus entre deux instants grâce à une semi-norme qui lui est propre.

Ces résultats sont également étendus aux processus sous-Gaussiens (section 2.2.1 van der Vaart & Wellner, 2000) où la majoration du sup est plus forte.

Définition A.2.2. *Un processus stochastique est dit* sous-Gaussien *pour la semi-métrique* d *si*

$$\mathbb{P}\left(|X_s - X_t| > x\right) \leq 2e^{\frac{1}{2}x^2/d^2(s,t)},\ \textit{pour touts}, t\in[0,T],\ x>0. \tag{A.14}$$

N'importe quel processus Gaussien est sous-Gaussien pour la semi-métrique de l'écart-type $d(s,t)=\sigma(X_s - X_t)$.

Corollaire A.2.2 (Corollary 2.2.8 p 101, van der Vaart & Wellner (2000)). *Soit* $\{X_t, t \in [0,T]\}$ *un processus sous-Gaussien séparable. Alors pour tout* $\delta > 0$*,*

$$E \sup_{d(s,t)\leq\delta} |X_s - X_t| \leq K \int_0^\delta \sqrt{logD(\epsilon, d)}\ d\epsilon, \tag{A.15}$$

pour une certaine constante K*. En particulier, pour tout* t_0*,*

$$E \sup_t |X_t| \leq E|X_{t_0}| + K \int_0^\infty \sqrt{logD(\epsilon, d)}\ d\epsilon. \tag{A.16}$$

Ce résultat permet d'avoir une majoration plus fine du sup du processus lorsque celui-ci est sous-Gaussien, et est utilisé pour prouver la consistance de nos estimateurs au chapitre 4.

Bibliographie

Billingsley, P. (1968). *Convergence of Probability Measures.* John Wiley, New York.

van der Vaart, A.W. & Wellner, J.A. (2000). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.

Bibliographie générale

[1] Adler, R. J. and Taylor, J. (2007). Random Fields and Geometry. Springer, New York.

[2] Baíllo, A. Cuevas, A. and Fraiman, R. (2010). Classification méthods for functional data. *The Oxford Handbook of Functional Data Analysis*, Edited by Ferraty, F. and Romain, Y., **10**, 259-297.

[3] Berger, Y. G. (1998). Rate of convergence for asymptotic variance of the Horvitz–Thompson estimator. *J. Statist. Planning and Inference*, **74**, 149-168.

[4] Besse, P.C. Cardot, H. and D. Stephenson (2000). Autoregressive Forecasting of Some Functional Climatic Variations. *Scandinavian Journal of Statistics*, Vol. **27**, 673-687.

[5] Bickel, P. J. & Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics*, **12**, 470-482.

[6] Billingsley, P. (1968). *Convergence of Probability Measures.* John Wiley, New York.

[7] Booth, J.G., Butler, R.W. and Hall, P. (1994). Bootstrap methods for finite population. *Journal of the American Statistical Association*, **89**, 1282-1289.

[8] Breidt, F. J. & Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, **28**, 1026-1053.

[9] Breidt, F. J. & Opsomer, J. D. (2008). Endogeneous post-stratification in surveys : classifying with a sample-fitted model. *Annals of Statistics*, **36**, 403-427.

[10] Bunea, F., Ivanescu, A. and M. Wegkamp (2011). Adaptive inference for the mean of a Gaussian process in functional data. *J. Roy. Statist. Soc. Ser. B*, to appear.

[11] Cardot, H., Ferraty, F. and P. Sarda (1999). Functional Linear Model. *Statistics and Probability Letters*. Vol. 45, **1**, 11-22.

[12] Cardot, H., Faivre, R. and M. Goulard (2003). Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data. *Journal of Applied Statistics*, Vol. **30**, 1185-1199.

[13] Cardot, H., Ferraty, F. and P. Sarda (2003). Spline Estimators for the Functional Linear Model. *Statistica Sinica*, Vol. 13, 571-591.

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

[15] Cardot, H., Dessertaine, A., Josserand, E. (2010b). Semiparametric models with functional responses in survey sampling setting : model assisted estimation of electricity consumption curves. *Compstat 2010*, Eds Lechevallier, Y. and Saporta, G. Physica-Verlag, Springer, 411-420.

[16] 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.

[17] Chauvet, G. (2007). Méthodes de Bootstrap en population finie. *PhD Thesis*, Université Rennes II, France.

[18] Chen, J. & Rao, J. N. K. (2007). Asymptotic normality under two-phase sampling designs. *Statistica Sinica*, **17**, 1047-1064.

[19] Chiky, R. (2009). Résumé de flux de données distribués. *PhD Thesis*, l'École Nationale Supérieure des Télécommunications, France.

[20] Chiky, R. & Hébrail, G. (2009). Spatio-temporal sampling of distributed data streams. *J. of Computing Science and Engineering*, to appear.

[21] Chiou, J.M., Müller, H.G. and Wang, J.L. (2003). Functional quasi-likelihood regression models with smooth random effects. *J.Roy. Statist. Soc., Ser. B*, **65**, 405-423.

[22] Claeskens, G. and van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Ann. Statist.*, **31**, 18521884.

[23] Cochran, W.G. (1977). *Sampling techniques.* 3rd Edition, Wiley, New York.

[24] 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.

[25] Degras, D. (2009). Nonparametric estimation of a trend based upon sampled continuous processes. *C. R. Math. Acad. Sci. Paris*, **347**, 191-194.

[26] Degras, D. (2010). Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica*, to appear.

[27] Dessertaine, A. (2008). Estimation de courbes de consommation électrique à partir de mesures non synchrones. In *Méthodes de sondage*, Eds. Guibert, P., Haziza, D., Ruiz-Gazen, A. and Tillé, Y. Dunod, Paris, 353-357.

[28] Deville J.C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE*, **15**, 3-101.

[29] Deville, J.C. & Tillé, Y. (2000). Balanced sampling by means of the cube method. Document de travail, Rennes, CREST-ENSAI.

[30] Dauxois J. & Pousse A. (1976). Les analyses factorielles en calcul des probabilités et en statistiques : essai d'étude synthétique. *Thèse d'état*, Université Paul Sabatier, Toulouse, France.

[31] Efron, B. (1979). Bootstrap methods : another look at the jackknife. *Annals of Statistics*, **7** ,1–26.

[32] Erdös, P. & Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hungar. Acad. Sci.* **4**, 49-61.

[33] Eubank, R.L. and Speckman P.L. (1993). Confidence Bands in Nonparametric Regression. *J. Amer. Statist. Assoc.*, **88**, 1287-1301.

[34] Faraway, J.T. (1997). Regression analysis for a functional response. *Technometrics*, **39**, 254-261.

[35] Ferraty, F. & Vieu, P. (2006). Nonparametric functional data analysis : theory and practice. *Springer Series in Statitics*, Springer.

[36] Ferraty, F. & Vieu, P. (2010). Kernel regression estimation for functional data. *The Oxford Handbook of Functional Data Analysis*, Edited by Ferraty, F. and Romain, Y., **4**, 72-129.

[37] Fey, A., van der Hofstad, R. and Klok, M. (2008). Large deviations for eigenvalues of sample covariance matrices, with applications to mobile communication systems. *Adv. in Appl. Probab.* **40**, 1048-1071.

[38] Fuller, W.A. (2009). *Sampling Statistics.* John Wiley and Sons.

[39] Geman, S. (1980). A Limit Theorem for the Norm of Random Matrices. *Ann. Probab.*, **8**, 252-261

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

[41] Hàjek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hungar. Acad. Sci.* **5**, 361-374.

[42] Hàjek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, **35**, 1491–1523.

Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hungar. Acad. Sci.* **5**, 361-374.

[43] Hall, P. (2010). Principal component analysis for functional data. *The Oxford Handbook of Functional Data Analysis*, Edited by Ferraty, F. and Romain, Y., **8**, 210-234.

[44] Hall, P., Müller, H.G. and Wang, J.L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.*, Vol 34, Number 3, 1493-1517.

[45] Hart, J. D. and Wehrly, T. E. (1993). Consistency of cross-validation when the data are curves. *Stoch. Proces. Applic.*, **45**, 351361.

[46] Helmers, R., Wegkamp, M. (1998). Wild Bootstrapping in Finite Population with Auxiliary Information. *Scandinavian journal of statistics*, **25**, 383-399.

[47] Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

[48] Isaki, C.T. & Fuller, W.A. (1982). Survey design under the regression superpopulation model. *J. Am. Statist. Ass.* **77**, 89-96.

[49] James, G. (2010). Sparseness and functional data analysis. *The Oxford Handbook of Functional Data Analysis*, Edited by Ferraty, F. and Romain, Y., **11**, 298-323.

[50] Johannes, J. (2008). Nonparametric Estimation in Functional Linear Model. *Functional and Operatorial Statistics*, Edited by Dabo-Niang, S. and Ferrraty, F., Physica-Verlag, **33**, 215-222.

[51] Kneip, A. and Utikal, K. J. (2001). Inference for Density Families Using Functional Principal Component Analysis. *Journal of the American Statistical Association*, **96**, 519–542.

[52] Krewski, D. and Rao, J. (1981). Inference from stratified samples : properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, **9**, 1010–1019.

[53] Krivobokova, T., Kneib, T. and G. Claeskens (2010). Simultaneous confidence bands for penalized spline estimators. *J. Am. Statist. Ass.*, **105**, 852-863.

[54] Landau, H. & Shepp, L.A. (1970). On the supremum of a Gaussian process. *Sankhyã*, **32**, 369-378

[55] Mas A. (2007). Testing for the mean of random curves : a penalization approach. *Statistical Inference for Stochastic Processes,* **10**, 147-163

[56] Müller, H.G. (2005). Functional modelling and classification of longitudinal data (with discussions). *Scand. J. Statist.*, **32**, 223-246.

[57] Müller, H.G. & Stadtmüller, U. (2005) Generalized functional linear models. *Ann. Statist.*, **33**, 774-805.

[58] Müller, H.G. Leng, X. (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, **22**, 68-76.

[59] Müller, H.G., Yao, F. (2008). Functional Additive Model. *J. Am. Statist. Ass.* **103**, 1534-1544.

[60] Neyman, J. (1934). On the two different aspects of representative method : the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, **97**, 558-606.

[61] Opsomer, J. D. and Miller, C. P. (2005). Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. *J. Nonparametric Statistics,* **17**, 593-611.

[62] Preda, C. & Saporta, G. (2002). Régression PLS sur une processus stochastique. *Revue de statistique appliquée*, **50**, 27-45.

[63] Ramsay, J. O. & Silverman, B.W. (2002). *Applied Functional Data Analysis : Methods and Case Studies*. Springer-Verlag.

[64] Ramsay, J. O. & Silverman, B.W. (2005). *Functional Data Analysis*. Springer-Verlag, 2nd ed.

[65] Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B*, **53**, 233-243.

[66] Rice, J. and Wu, C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253-259.

[67] Robinson, P. M. & Särndal, C. E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya : The Indian Journal of Statistics*, **45**, 240-248.

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

[69] Sood, A., James, G. and Tellis, G. (2009). Functional Regression : A New Model for Predicting Market Penetration of New Products. *Marketing Science*, **28**, 36-51.

[70] Sun, J. and Loader, C.R. (1994). Simultaneous confidence bands for linear regression and smoothing. *Annals of Statistics*, **22**, 1328-1345.

[71] Särndal, C. E., Swensson, B. and J. Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

[72] Tillé, Y. (2001). Théorie des sondages : Échantillonnage et estimation en populations finies. *Dunod*, Paris.

[73] Tillé, Y. (2006). Sampling Algorithms. *Springer*, New York.

[74] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation.* Springer, New York.

[75] Van der Vaart, A.W. & Wellner, J.A. (2000). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

[76] Yao, F. Müeller, H.G. and Wang, J.L. (2005). Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, 577-590.

[77] Yates, F. and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, *B15*, 235–261.

[78] Zhang, J. T. and Chen, J. (2007). Statistical inferences for functional data. *Annals of Statistics*, **35**, 1052-1079.