HAL
open science

Estimation de synchrones de consommation électrique par sondage et prise en compte d'information auxiliaire

Pauline Lardin

To cite this version:

Pauline Lardin. Estimation de synchrones de consommation électrique par sondage et prise en compte d'information auxiliaire. Mathématiques générales [math.GM]. Université de Bourgogne, 2012. Français. NNT : 2012DIJOS049 . tel-00842199

HAL Id: tel-00842199

https://theses.hal.science/tel-00842199v1

Submitted on 8 Jul 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

par

Pauline Lardin

Estimation de synchrones de consommation électrique par sondage et prise en compte d'information auxiliaire

Soutenue le 26 novembre 2012 devant le jury composé de :

David Haziza	Université Montréal et ENSAI	Rapporteur
Pascal Sarda	Université Paul Sabatier Toulouse	Rapporteur
Hervé Cardot	Université de Bourgogne	Directeur de thèse
Camelia Goga	Université de Bourgogne	Directrice de thèse
Guillaume Chauvet	CREST - ENSAI	Examinateur
Anne Ruiz-Gazen	Université Toulouse Capitole	Examinatrice
Jérôme Cubillé	EDF R&D - ICAME	Invité
Alain Dessertaine	La poste - Pôle Expertise	Invité

Remerciement

Je tiens à remercier mes directeurs de thèse Hervé Cardot et Camelia Goga ainsi que Alain Dessertaine de m'avoir donné l'opportunité de faire cette thèse ainsi que pour leur disponibilité, leurs diverses idées et la confiance qu'ils m'ont accordée tout au long de ces trois ans. Ils m'ont transmis de nombreuses connaissances en théorie des sondages et ils m'ont montré qu'il est tout à fait possible d'allier le monde de la recherche à celui de l'entreprise.

Je souhaite également remercier David Haziza et Pascal Sarda d'avoir consacré du temps à la lecture de mon travail ainsi qu'à l'écriture des rapports. Je remercie Guillaume Chauvet, Jérôme Cubillé et Anne Ruiz-Gazen d'avoir accepté d'être dans le jury.

Je remercie l'ensemble des équipes SOAD et E74 à EDF dont Alina, Leslie, Silvia, Sophie et Pierre pour l'ambiance très agréable qu'ils ont su maintenir au quotidien mais aussi pour leur disponibilité, les divers échanges, professionnels ou non, que nous avons eu au long de cette thèse. Ceux-ci m'ont permis de mieux appréhender les données et les problématiques de mon sujet. Je remercie également EDF dont le soutien administratif et financier a rendu cette thèse possible.

Je remercie tous les membres de l'institut de Mathématiques de Bourgogne et plus particulièrement l'équipe SPAN et Etienne Josserand. La qualité de leurs cours et l'émulation en émanant m'ont orienté vers cette thèse avec succès.

Je voudrais remercier ma famille et plus particulièrement mes parents qui m'ont toujours encouragée à poursuivre mes études et ce dans les meilleures conditions possibles. Enfin, je remercie tout spécialement Loïc pour son soutien dans les moments difficiles et ses encouragements tout au long de ces trois années.

Résumé

Dans cette thèse, nous nous intéressons à l'estimation de la synchrone de consommation électrique (courbe moyenne). Etant donné que les variables étudiées sont fonctionnelles et que les capacités de stockage sont limitées et les coûts de transmission élevés, nous nous sommes intéressés à des méthodes d'estimation par sondage, alternatives intéressantes aux techniques de compression du signal. Nous étendons au cadre fonctionnel des méthodes d'estimation qui prennent en compte l'information auxiliaire disponible afin d'améliorer la précision de l'estimateur de Horvitz-Thompson de la courbe moyenne de consommation électrique. La première méthode fait intervenir l'information auxiliaire au niveau de l'estimation, la courbe moyenne est estimée à l'aide d'un estimateur basé sur un modèle de régression fonctionnelle. La deuxième l'utilise au niveau du plan de sondage, nous utilisons un plan à probabilités inégales à forte entropie puis l'estimateur de Horvitz-Thompson fonctionnel. Une estimation de la fonction de covariance est donnée par l'extension au cadre fonctionnel de l'approximation de la covariance donnée par Hájek. Nous justifions de manière rigoureuse leur utilisation par une étude asymptotique. Pour chacune de ces méthodes, nous donnons, sous de faibles hypothèses sur les probabilités d'inclusion et sur la régularité des trajectoires, les propriétés de convergence de l'estimateur de la courbe moyenne ainsi que de sa fonction de covariance. Nous établissons également un théorème central limite fonctionnel.

Afin de contrôler la qualité de nos estimateurs, nous comparons deux méthodes de construction de bande de confiance sur un jeu de données de courbes de charge réelles. La première repose sur la simulation de processus gaussiens. Une justification asymptotique de cette méthode sera donnée pour chacun des estimateurs proposés. La deuxième utilise des techniques de bootstrap qui ont été adaptées afin de tenir compte du caractère fonctionnel des données.

Mots clés : Approximation de Hájek, bande de confiance, bootstrap, données fonctionnelles, estimateur de Horvitz-Thompson, estimateur model-assisted, fonction de covariance, modèle linéaire fonctionnel, plan à probabilités inégales sans remise, théorème central limite fonctionnel, sondage.

Abstract

In this thesis, we are interested in estimating the mean electricity consumption curve. Since the study variable is functional and storage capacities are limited or transmission cost are high survey sampling techniques are interesting alternatives to signal compression techniques. We extend, in this functional framework, estimation methods that take into account available auxiliary information and that can improve the accuracy of the Horvitz-Thompson estimator of the mean trajectory. The first approach uses the auxiliary information at the estimation stage, the mean curve is estimated using model-assisted estimators with functional linear regression models. The second method involves the auxiliary information at the sampling stage, considering πps (unequal probability) sampling designs and the functional Horvitz-Thompson estimator. Under conditions on the entropy of the sampling design the covariance function of the Horvitz-Thompson estimator can be estimated with the Hájek approximation extended to the functional framework. For each method, we show, under weak hypotheses on the sampling design and the regularity of the trajectories, some asymptotic properties of the estimator of the mean curve and of its covariance function. We also establish a functional central limit theorem.

Next, we compare two methods that can be used to build confidence bands. The first one is based on simulations of Gaussian processes and is assessed rigorously. The second one uses bootstrap techniques in a finite population framework which have been adapted to take into account the functional nature of the data.

Keywords : Bootstrap, confidence band, covariance function, functional central limit theorem, functional data, functional linear model, Hájek variance approximation, Horvitz-Thompson estimator, model-assisted estimator, survey sampling, unequal probability sampling without replacement.

Table des matières

Introduction **14**

1 Quelques rappels sur les sondages **19**
1.1 Notations 19
1.2 Estimateur de Horvitz-Thompson 21
1.3 Prise en compte de l'information auxiliaire 23
1.3.1 Au niveau du tirage de l'échantillon 24
1.3.2 Au niveau de l'estimation 30
1.4 Présentation de l'approche modèle 32
1.5 Asymptotique en théorie des sondages 38
1.6 Calcul de la précision 40
1.6.1 La technique de linéarisation 40
1.6.2 Le Jackknife 41
1.6.3 Le bootstrap 42
1.7 Sondage sur données fonctionnelles 43
1.7.1 Estimateur de Horvitz-Thompson pour données fonctionnelles . 43
1.7.2 Estimation à partir de trajectoires discrétisées 45
1.7.3 Quelques rappels de probabilité 46

2 Estimation d'une trajectoire moyenne par l'approche model-assisted **49**
2.1 Estimation à l'aide d'un modèle linéaire fonctionnel 49
2.1.1 Un estimateur régularisé pour le cadre asymptotique 51
2.1.2 Estimation à l'aide de trajectoires discrétisées 52
2.2 Propriétés asymptotiques sous le plan de sondage 53
2.2.1 Hypothèses 53
2.2.2 Convergence uniforme 54
2.2.3 Estimation de la fonction de covariance sous le plan de sondage 55
2.3 Normalité asymptotique 56
2.4 Preuves 57
2.4.1 Quelques lemmes utiles 57
2.4.2 Preuve de la Proposition 2.1 et de la Proposition 2.2 63
2.4.3 Preuve de la convergence de la fonction de covariance 67
2.4.4 Preuve de la normalité asymptotique 72

3 Estimation de la variance pour des plans à probabilités inégales à forte entropie 75
3.1 Estimation de la covariance pour les plans à fortes entropies 76
3.2 Propriétés asymptotiques 77
3.2.1 Hypothèses 77
3.2.2 Convergence et propriétés asymptotiques 78
3.3 Exemple : estimation de la variance de la courbe de consommation électrique 80
3.4 Preuves 86
3.4.1 Quelques lemmes utiles 86
3.4.2 Preuve de la proposition 3.3 88
3.4.3 Preuve de la proposition 3.4 88
3.4.4 Preuve de la Proposition 3.5 94

4 Construction de bandes de confiance 95
4.1 Introduction 95
4.2 Par simulation de processus gaussien 97
4.3 Par bootstrap 99
4.4 Etude de la courbe de consommation moyenne d'électricité 106
4.4.1 Description des stratégies utilisées 106
4.4.2 Erreur d'estimation de la courbe moyenne 108
4.4.3 Taux de couverture et largeur des bandes de confiance 109
4.4.4 Temps de calcul 112
4.5 Conclusion 113

Conclusion et perspectives 115

Annexe 120

A Prise en compte de la température 123
A.1 Estimateur assisté par un modèle basé sur l'ACP fonctionnelle 124
A.2 Estimation de la courbe moyenne de consommation des particuliers . . 125
A.2.1 Recherche du modèle de superpopulation 126
A.2.2 Mise en place du plan de sondage 131
A.3 Conclusion des travaux 135

Bibliographie 137

Table des figures

1.1 Représentation de la consommation à un instant t en fonction de la consommation moyenne de la semaine précédente. 27
1.2 Echantillon de 10 courbes de consommation électrique. La courbe moyenne est tracée en gras. 44

3.1 Courbe moyenne de la consommation sur la population et son estimation obtenue à l'aide de l'échantillon s' de taille $n = 1500$ 82
3.2 Variance empirique γ_{emp}, approximation de Hájek γ_H et la variance estimée $\hat{\gamma}_{H,d}$ obtenue à l'aide de l'échantillon s' de taille $n = 1500$ . . . 83
3.3 Erreur d'estimation $\gamma_{emp}(t,r) - \hat{\gamma}_{H,d}(t,r)$ obtenue à l'aide de l'échantillon s' de taille $n = 1500$ 84
3.4 Erreur d'approximation $\gamma_{emp}(t,r) - \gamma_H(t,r)$ pour $n = 1500$ 85
3.5 Représentation du logarithme des poids de sondage ($\log(1/\pi_k)$) pour les différentes tailles d'échantillon 86

4.1 Représentation de la consommation à un instant t en fonction de la consommation moyenne de la semaine précédente. 106
4.2 Exemples de bande de confiance 110
4.3 Sondage aléatoire simple sans remise. Largeur des bandes de confiance ponctuelles, globales par simulations de processus et avec Bonferroni ($\alpha = 0.05$). 112
4.4 Sondage stratifié (STRAT 1). Largeur des bandes de confiance ponctuelles, globales par simulations de processus et avec Bonferroni (avec $\alpha = 0.05$). 113

A.1 Sortie de l'ACP fonctionnelle du jour J 126
A.2 Représentation graphique entre la première composante et (a) la consommation moyenne de la semaine précédente (b) la température moyenne du jour J. 127
A.3 Représentation graphique entre la deuxième composante et (a) la consommation moyenne de la semaine précédente (b) la température moyenne du jour J. 128
A.4 Thermosensibles : Régression de la consommation à l'instant où la température est la plus corrélée avec la consommation 129
A.5 Thermosensibles : Régression du log de la consommation à l'instant où la température est la plus corrélée avec la consommation 129

A.6 Thermosensibles : Régression du log de la consommation à l'instant où la température est la moins corrélée avec la consommation (sans Heure Creuse) 130
A.7 Thermosensibles : Régression du log de la consommation à l'instant où la température est la moins corrélée avec la consommation (avec Heure Creuse) 130
A.8 Non-Thermosensibles : Régression à l'instant où la température est la plus corrélée avec la consommation 130
A.9 Comparaison des erreurs moyennes d'estimation des clients thermosensibles 134
A.10 Comparaison des erreurs moyennes d'estimation des clients non-thermosensibles 134

Liste des tableaux

3.1 $RMSE(\hat{\gamma}_{H,d})$, $RB(\hat{\gamma}_{H,d})^2$ et l'erreur d'estimation $R_2(\hat{\gamma}_{H,d})$ pour différentes tailles d'échantillon, avec I = 10000 échantillons. 82

4.1 STRAT 1 : stratification à partir des courbes. Les strates sont construites à partir des courbes de la semaine 1. L'allocation n_h optimale est calculée à partir des courbes de la semaine 1. 107
4.2 STRAT 2 : stratification à partir de la consommation moyenne x_k. L'allocation optimale n_h est calculée à partir de la consommation moyenne de la semaine 1. 108
4.3 Erreur R_1 d'estimation de la moyenne μ, avec I = 10000 réplications. . 108
4.4 Erreur quadratique R_2 d'estimation de la moyenne μ, avec I = 10000 réplications. 109
4.5 Taux de couverture empirique (en %), pour I=2000 réplications. 111
4.6 Largeur moyenne des bandes de confiance, pour I = 2000 réplications. . 111
4.7 Temps d'exécution d'une simulation en secondes pour M=5000 réplications. Les stratégies SRSWOR, MA et STRAT ont été programmés avec R et πps avec SAS. 113

A.1 Erreur moyenne des clients thermosensibles 133
A.2 Erreur moyenne des clients non-thermosensibles 133

Introduction

Afin de mieux intégrer les énergies renouvelables intermittentes décentralisées et d'améliorer la gestion du réseau de distribution, l'ensemble des compteurs électriques vont être remplacés dans les années à venir par des compteurs communicants. Ceux-ci permettrons également aux clients d'être acteurs de leur propre consommation. Electricité de France (EDF) aura alors la possibilité de mesurer la courbe de consommation électrique de chacun de ses clients à des pas potentiellement très fins (10 minutes, 1/2 heure, etc.). Cependant la mise en place de cette installation soulève de nouvelles problématiques informatiques et statistiques : comment rapatrier, stocker et analyser de très grandes bases de données concernant des phénomènes qui évoluent au cours du temps ? Des travaux de recherches sont actuellement en cours à EDF afin d'élaborer un système d'information permettant de gérer et d'utiliser au mieux ces informations (techniques de résumé et d'analyse de flux de données, techniques d'échantillonnage et d'estimation développées dans le cadre de la théorie des sondages). Dans cette thèse, nous allons plus particulièrement nous intéresser à l'estimation de la courbe moyenne de la consommation électrique par sondage. Les pas de discrétisation étant très fins, les unités statistiques étudiées peuvent être considérées comme des fonctions du temps. Nous pouvons alors faire appel à des outils d'analyse de données fonctionnelles pour décrire les données et construire des modèles statistiques. Ces outils sont apparus dans les années 1970 (Deville (1974), Dauxois et Pousse (1976)) mais se sont réellement développés au cours des années 1990 avec l'augmentation des performances des ordinateurs et des capacités de stockage. Les applications concernent différents domaines tels que l'économie, la médecine ou encore la télédétection. Dans leur livre Ramsay et Silverman (2005) présentent un panorama des différentes techniques d'analyse fonctionnelle : analyse en composantes principales fonctionnelle, modèle de régression pour des variables explicatives et/ou expliquées fonctionnelles, etc. Nous trouvons également dans cet ouvrage des méthodes de lissage par splines, noyaux ou polynômes locaux qui permettent de prendre en compte le schéma de discrétisation dans les méthodes d'analyse fonctionnelle. Dans leur ouvrage, Ferraty et Vieu (2006) proposent également des méthodes non-paramétriques.

Avec la mise en place des nouveaux compteurs, nous serons amenés à travailler sur une base de données très grande : 140 Téra Octet seront nécessaires pour stocker la consommation au pas 10 minutes pendant un an des 35 millions d'individus. Il sera alors impossible de collecter, stocker et analyser l'ensemble de ces données. Chiky (2009) a montré que lorsqu'on s'intéresse à des indicateurs simples, tels que la moyenne, les techniques de sondages sont une alternative intéressante aux techniques

de compression du signal et qu'elles permettent d'obtenir des estimations précises à un coût raisonnable.

Les travaux reliant analyse de données fonctionnelles et théorie des sondages sont encore peu nombreux dans la littérature statistique. L'estimateur de Horvitz-Thompson d'une courbe moyenne a été introduit par Cardot *et al.* (2010a) et ses propriétés de convergence uniforme ont été démontrées par Cardot et Josserand (2011). Lorsque le théorème central limite est satisfait et que l'on dispose d'un estimateur précis de la fonction de covariance, Cardot *et al.* (2012) ont donné une justification asymptotique d'une méthode de construction de bandes de confiance. Ils déterminent à l'aide de simulations la loi du supremum d'un processus gaussien centré dont la fonction de covariance est la fonction de covariance estimée sur l'échantillon. Certaines méthodes ont également été développées pour prendre en compte l'information auxiliaire disponible et ainsi améliorer la précision de l'estimateur de Horvitz-Thompson. Cardot et Josserand (2011) ont estimé la courbe de consommation à partir d'un sondage stratifié à allocation optimale fonctionnelle. Cardot *et al.* (2010b) proposent également une approche modèle assisté semi-paramétrique basée sur l'analyse en composante principale fonctionnelle. L'estimation de la médiane fonctionnelle par sondage a été étudiée par Chaouch et Goga (2012).

L'objectif de la présente thèse est d'étendre au cadre fonctionnel des méthodes d'estimation qui prennent en compte l'information auxiliaire disponible pour améliorer la précision de l'estimateur de la courbe moyenne de consommation et de justifier de manière rigoureuse leur utilisation par une étude asymptotique. La première méthode mise en place est basée sur un modèle de régression fonctionnelle et la seconde utilise un plan à probabilités inégales à forte entropie et l'estimateur de Horvitz-Thompson fonctionnel. Afin de contrôler la qualité de ces estimations, nous avons comparé deux méthodes de construction de bandes de confiance sur un jeu de données de courbes de charge de EDF. La première a été proposée par Cardot *et al.* (2012) et elle est basée sur la simulation de processus gaussien. La deuxième fait intervenir des techniques de Bootstrap (Booth *et al.* (1994), Chauvet (2007)) qui ont été adaptées afin de tenir compte du caractère fonctionnel des données. Ces méthodes pourront par la suite être utilisées pour améliorer la connaissance par usage (chauffage, eau chaud sanitaire, voiture électrique, etc.) ou par segment de client (type de tarif, résidentiels, industriels) afin de proposer de nouvelles offres tarifaires mieux adaptées aux nouveaux usages. Elles permettront enfin de mieux équilibrer l'offre et la demande en électricité.

Le chapitre 1 présente quelques rappels sur la théorie des sondages en population finie. Nous décrivons quelques méthodes qui permettent de prendre en compte les variables auxiliaires pour améliorer la précision de l'estimateur de Horvitz-Thompson : tirage à probabilités inégales, post-stratification, calage, estimateur assisté par un modèle. Nous introduisons également la notion d'asymptotique en théorie des sondages. Nous présentons également des méthodes qui permettent d'estimer la précision de l'estimateur sans utiliser une formule analytique de la variance. Pour finir, nous introduisons l'estimateur de Horvitz-Thompson d'une courbe moyenne et nous définissons

les notions de convergence dans l'espace fonctionnel.

Dans le chapitre 2, nous proposons d'estimer la courbe moyenne à l'aide d'un estimateur basé sur un modèle de régression fonctionnelle (Faraway (1997), Ramsay et Silverman (2005)). Celui-ci peut être vu comme une extension directe au cadre fonctionnel de l'estimateur par la régression généralisée (GREG) étudié par Robinson et Särndal (1983) et Särndal *et al.* (1992). Sous des hypothèses classiques sur le plan de sondage et sur la régularité des trajectoires, nous montrons la convergence uniforme de notre estimateur ainsi que de celle de sa fonction de variance estimée. Enfin, nous démontrons le théorème central limite fonctionnel.

Dans le chapitre 3, nous proposons, pour les plans à forte entropie, d'estimer la fonction de covariance de l'estimateur de Horvitz-Thompson par l'extension au cadre fonctionnel de l'approximation de Hájek (1964). L'intérêt de cette approximation est qu'elle ne fait intervenir que les probabilités d'inclusion d'ordre un. Nous montrons que, sous certaines hypothèses sur les probabilités d'inclusion et sur la régularité des trajectoires, l'estimateur de la fonction de variance converge uniformément vers la fonction de variance de Horvitz-Thompson. Dans le cas d'un tirage réjectif, nous donnons également sa vitesse de convergence. Enfin nous appliquons cette méthode d'estimation sur un jeu de données de courbes de consommation.

Dans le chapitre 4, nous présentons deux algorithmes de construction de bande de confiance. Le premier est basé sur la simulation de processus gaussien et s'inspire de techniques basées sur l'estimation de la fonction de covariance de l'estimateur (Cardot *et al.* (2012)). Une justification asymptotique pour les plans à forte entropie et pour l'estimation basée sur un modèle de régression fonctionnelle est donnée. Le deuxième repose sur des techniques de Bootstrap adaptées aux populations finies. Une comparaison des différentes stratégies est faite à partir de l'estimation des courbes de charge d'EDF.

Dans l'annexe A, nous essayons d'apporter un début de réponse à la question suivante : Est-ce que l'utilisation de variables auxiliaires telles que la température extérieure permet d'améliorer la précision de l'estimateur de la courbe moyenne de consommation électrique des particuliers ? Une description du modèle d'estimation mis en place sera donnée. Nous comparerons ensuite sa précision à celles de plans plus classiques. Pour des raisons de confidentialité certaines parties de l'étude ne sont pas présentées dans ce document.

Ces différents travaux ont fait l'objet de plusieurs rapports techniques

- H. Cardot, A. Dessertaine, C. Goga, E. Josserand et P. Lardin. Comparaison de différents plans de sondage et construction de bandes de confiance pour l'estimation de la moyenne de données fonctionnelles : une illustration sur la consommation électrique. A paraitre dans *Survey Methodology* et disponible sur http ://hal.archives-ouvertes.fr/docs/00/71/47/29/PDF/Rev1fCDGJL2012.pdf.
- H. Cardot, C. Goga et P. Lardin. Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. En révision dans *Electronic Journal of Statistics* et disponible sur arXiv http ://arxiv.org/pdf/1204.6382.pdf.
- H. Cardot, C. Goga et P. Lardin. Variance estimation and asymptotic confidence bands for the mean estimator of sampled functional data with high entropy unequal probability sampling designs. Soumis à *Scandinavian Journal of Statistics* et disponible sur http ://arxiv.org/pdf/1209.6503v2.pdf.

et de présentations à des conférences

- P. Lardin, H. Cardot, A. Dessertaine, C. Goga et E. Josserand. Estimation and confidence bands for the mean electricity consumption curve : a comparison of unequal probability sampling designs and model-assisted approaches. *ISI, Dublin, 2011.*
- P. Lardin, H. Cardot et C. Goga. Convergence uniforme de l'estimateur d'une trajectoire moyenne assistée par un modèle de régression linéaire fonctionnelle. *Journée de Statistiques, Bruxelles, 2012.*
- P. Lardin, H. Cardot et C. Goga. Théorème centrale limite et bandes de confiance asymptotiques pour l'estimation de la moyenne de données fonctionnelles pour des plans à probabilités inégales. *Colloque francophone sur les sondages, Rennes, 2012.*

Chapitre 1

Quelques rappels sur les sondages

Dans ce chapitre, nous rappelons quelques éléments de la théorie des sondages qui nous seront utiles dans les chapitres suivants. Nous présentons quelques méthodes d'estimation qui utilisent l'information auxiliaire afin d'améliorer la précision de l'estimateur de Horvitz-Thompson : tirage à probabilités inégales, post-stratification, calage, estimateur assisté par un modèle. Nous introduisons également la notion d'asymptotique en théorie des sondages. Nous présentons ensuite des méthodes qui permettent d'estimer la précision de l'estimateur sans utiliser une formule analytique de la variance. Pour de plus amples détails, on pourra se référer aux ouvrages suivants : Särndal *et al.* (1992), Tillé (2001) et Ardilly (2006). Pour finir, nous introduisons l'estimateur de Horvitz-Thompson d'une courbe moyenne et nous définissons les notions de convergence dans l'espace fonctionnel.

1.1 Notations

Considérons une population finie U composée de N éléments

$$U = \{u_1, ..., u_k, ..., u_N\} = \{1, ..., k, ..., N\},$$

où pour simplifier, nous identifions le k-ème élément de U noté u_k par k. Dans la suite, nous supposons que les individus k peuvent être identifiés sans ambiguïté.

Soit Y une variable d'intérêt définie pour chaque individu k de la population U. On notera Y_k la valeur de Y prise par l'individu k. L'objectif est d'estimer une fonction de la variable d'intérêt

$$\theta = \theta(Y_k, k \in U)$$

que l'on appellera *paramètre d'intérêt ou fonctionnelle* et d'évaluer la précision de cette estimation. Parmi les fonctionnelles les plus couramment étudiées nous retrouvons :

– le total

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

– la moyenne

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

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

On peut également s'intéresser à des fonctionnelles plus complexes, telles que
– le ratio des totaux de deux variables Y et Z
$$R = \frac{t_Y}{t_Z},$$
– le fractile f_α d'ordre α de la variable Y
$$f_\alpha = \mathrm{Inf}\{x; F(x) \geq \alpha\} \text{ où } F(x) = \tfrac{1}{N}\textstyle\sum_{k\in U} \mathbb{1}_{\{Y_k \leq x\}},$$
avec $\mathbb{1}_{\{Y_k\leq x\}} = 1$ si $Y_k \leq x$, et 0 sinon.

Si nous effectuons un recensement de la population, il est théoriquement possible d'obtenir les valeurs des différentes fonctionnelles. Cependant dans la pratique, il est difficile de mettre en place un recensement et ce pour différentes raisons :
– Ce type d'opération est très coûteux (mise en place, collecte de l'information, exploitation des données, etc.).
– Cela ne permet pas d'obtenir des estimations rapides des fonctionnelles.

On se contente donc généralement de faire une enquête par sondage. Cela permet, notamment, d'obtenir des estimations beaucoup plus rapidement (moins de temps de saisie et de contrôle), de diminuer les coûts et d'approfondir certains domaines qui ne peuvent être qu'effleurés lors des recensements (cf. Ardilly (2006), chapitre 1).

Nous supposerons que l'échantillon s de taille n_s est tiré selon un plan de sondage $p(s)$ sans remise, où $p(.)$ est une loi de probabilité sur l'ensemble des parties $\mathcal{S}$ de U. Le plan de sondage p vérifie
$$\forall s \subset \mathcal{S} \quad p(s) \geq 0 \text{ et } \sum_{s\subset\mathcal{S}} p(s) = 1.$$

Le plan de sondage p sera dit *de taille fixe* si la taille de l'échantillon est fixée. Dans ce cas, on notera n la taille de l'échantillon. Par la suite nous étudierons principalement des plans de sondage à taille fixe.

On notera $\mathbb{E}_p(.)$ (respectivement $V_p(.)$) l'espérance (respectivement la variance) sous le plan de sondage p. Pour un estimateur $\hat{\theta}$ de θ, on a
$$\mathbb{E}_p(\hat{\theta}) = \sum_{s\subset\mathcal{S}} p(s)\hat{\theta}(s)$$
et
$$V_p(\hat{\theta}) = \sum_{s\subset\mathcal{S}} p(s)\left(\hat{\theta}(s) - \mathbb{E}_p(\hat{\theta})\right)^2.$$

1.2 Estimateur de Horvitz-Thompson

Notons $\mathbb{1}_k = \mathbb{1}_{k \in s}$ l'indicatrice d'appartenance de l'individu k à l'échantillon s.

Pour un plan de sondage $p(.)$, on appelle *probabilité d'inclusion d'ordre un* la probabilité π_k de l'individu k d'être sélectionné dans un échantillon. Pour tout $k \in U$,

$$\pi_k = \Pr(k \in s) = \sum_{s \ni k} p(s).$$

On appelle *probabilité d'inclusion d'ordre deux* la probabilité π_{kl} que deux unités distinctes k et l soient sélectionnées dans un échantillon. Pour tous $k, l \in U$

$$\pi_{kl} = \Pr(k \in s \text{ et } l \in s) = \sum_{s \ni k \& l} p(s).$$

Par la suite, nous allons supposer que les probabilités d'inclusion π_k et π_{kl} sont strictement positives pour tout $k \neq l \in U$.

Nous pouvons facilement montrer les résultats suivants (cf. Tillé (2001), chapitre 3).

Résultat 1.1. *Pour un plan de sondage $p(.)$ fixé, la fonction $\mathbb{1}_k$ a les propriétés suivantes :*

- $\mathbb{E}_p(\mathbb{1}_k) = \pi_k$;
- $V_p(\mathbb{1}_k) = \pi_k(1 - \pi_k) = \Delta_{kk}$;
- $Cov_p(\mathbb{1}_k, \mathbb{1}_l) = \pi_{kl} - \pi_k \pi_l = \Delta_{kl}$, $k \neq l$ *pour tous* $k, l \in U$.

Résultat 1.2. *Soit p un plan de taille fixe égale à n. Alors :*

- $\sum_U \pi_k = n$;
- $\sum_{k \neq l} \sum_U \pi_{kl} = n(n-1)$;
- $\sum_{k \neq l, l \in U} \pi_{kl} = (n-1)\pi_k$.

Théorème 1.1. *Horvitz et Thompson (1952)*

Si pour toute unité $k \in U$ on a $\pi_k > 0$, alors

$$\hat{t}_Y = \sum_{k \in s} \frac{Y_k}{\pi_k} = \sum_{k \in U} \frac{Y_k}{\pi_k} \mathbb{1}_k$$

est un estimateur sans biais de t_Y.

Cet estimateur est appelé *estimateur de Horvitz-Thompson* ou *π-estimateur du total t_Y*. On utilise également le terme estimateur par les valeurs dilatées puisqu'il s'agit d'un estimateur pondéré qui affecte un poids $d_k = 1/\pi_k \geq 1$ à chaque unité k de l'échantillon. Ainsi, chaque unité k de l'échantillon représente $1/\pi_k$ unités de la population dans l'estimation du total.

Remarque 1.1.

1. *L'estimateur de Horvitz-Thompson est le seul estimateur linéaire sans biais d'un total de la forme*
$$\sum_{k\in s} w_k Y_k$$
, où les w_k ne dépendent pas de l'échantillon (cf. Ardilly (2006), chapitre 2).
2. *La partie aléatoire de l'estimateur repose uniquement sur la sélection ou la non-sélection d'un individu dans l'échantillon.*

La variance de l'estimateur de Horvitz-Thompson est donnée par le théorème suivant :

Théorème 1.2. *Horvitz et Thompson (1952)*
L'estimateur de Horvitz-Thompson $\hat{t}_Y$ a pour variance
$$V_p(\hat{t}_Y) = \sum_{k\in U}\sum_{l\in U} \frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}(\pi_{kl} - \pi_k\pi_l).$$
Si $\pi_{kl} > 0$, $\forall k, l \in U$, alors un estimateur non biaisé de cette variance est donné par
$$\hat{V}_p(\hat{t}_Y) = \sum_{k\in s}\sum_{l\in s} \frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}.$$

Lorsque le plan est de taille fixe, Sen (1953) et Yates et Grundy (1953) ont montré qu'il est possible de reformuler la variance.

Théorème 1.3. *Sen (1953) ; Yates et Grundy (1953)*
Si le plan est de taille fixe, l'estimateur de Horvitz-Thompson $\hat{t}_Y$ a pour variance
$$V_p(\hat{t}_Y) = -\frac{1}{2}\sum_{k\in U}\sum_{l\neq k\in U} \Delta_{kl}\left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2, \tag{1.1}$$
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
$$\hat{V}_p(\hat{t}_Y) = -\frac{1}{2}\sum_{k\in s}\sum_{l\neq k\in s} \frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2 \tag{1.2}$$
si $\pi_{kl} > 0$ $\forall k, l \in U$. Cet estimateur est appelé estimateur de Sen-Yates-Grundy.

Une condition suffisante pour que cet estimateur soit positif est que les conditions de Sen-Yates-Grundy soient satisfaites, *i.e.*
$$\pi_k\pi_l - \pi_{kl} \geq 0 \quad \forall k \neq l \in U.$$

Nous verrons dans le chapitre 3 que ces conditions sont satisfaites asymptotiquement pour les plans à forte entropie.

Exemple : Sondage aléatoire simple sans remise, noté SRSWOR

Le sondage aléatoire simple sans remise est très utilisé en pratique et a la propriété de ne nécessiter aucune information auxiliaire sur les individus lors de sa mise en œuvre. Il est à la base de plans de sondage plus complexes tels que le sondage stratifié et le sondage à plusieurs degrés.

Le sondage aléatoire simple sans remise consiste à tirer un échantillon de taille n parmi les N individus de la population sans faire intervenir d'information auxiliaire et sans manipulation préalable sur les individus de la population. Un tirage aléatoire simple sans remise attribue à tous les échantillons de même taille n fixée, la même probabilité d'être tiré. Ainsi un échantillon s de taille n a une probabilité de sélection $p(s)$ égale à l'inverse du nombre d'échantillons distincts de taille n que l'on peut constituer à partir d'une population de taille N :

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

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

$$\begin{aligned} \pi_k &= \frac{C_{N-1}^{n-1}}{C_N^n} = \frac{n}{N}, \quad k \in U, \\ \pi_{kl} &= \frac{C_{N-2}^{n-2}}{C_N^n} = \frac{n(n-1)}{N(N-1)}, \quad k \neq l \in U. \end{aligned}$$

L'estimateur de Horvitz-Thompson $\hat{\mu}_Y$ de la moyenne μ_Y devient simplement

$$\hat{\mu}_Y = \frac{1}{n} \sum_{k \in s} Y_k \tag{1.3}$$

et sa variance est donnée par

$$V(\hat{\mu}_Y) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{Y,U}^2 \tag{1.4}$$

avec $S_{Y,U}^2 = \frac{1}{N-1} \sum_U (Y_k - \mu_Y)^2$.

Celle-ci peut être estimée par

$$\hat{V}(\hat{\mu}_Y) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{Y,s}^2$$

avec $S_{Y,s}^2 = \frac{1}{n-1} \sum_s (Y_k - \hat{\mu}_Y)^2$.

1.3 Prise en compte de l'information auxiliaire

Il est bien connu que la consommation électrique des individus est fortement dépendante de variables telles que leurs consommations passées, les caractéristiques géographiques ou météorologiques. Celles-ci peuvent être prises en compte dans le plan

de sondage. Une première possibilité est d'utiliser ces variables auxiliaires au niveau du tirage de l'échantillon : sondage stratifié, sondage à probabilités inégales, sondage équilibré, etc. Une autre possibilité est de faire intervenir cette information au niveau de l'estimation : post-stratification, calage, approche assistée par un modèle, etc. Dans les sections suivantes, nous allons détailler quelques unes de ces méthodes. Pour plus de détails, on pourra se référer à Ardilly (2006), Tillé (2001) et Särndal *et al.* (1992).

1.3.1 Au niveau du tirage de l'échantillon

Le sondage stratifié, noté STRAT

Nous supposons que la population U est partitionnée en H strates, $U_1, ..., U_H$, de taille $N_1, ..., N_H$ connues. Les strates vérifient donc

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

et

$$N_1 + ... + N_H = N.$$

Pour constituer notre échantillon, nous prélevons indépendamment dans chaque strate U_h un échantillon s_h de taille n_h, selon un plan de sondage quelconque. Ainsi l'estimateur de Horvitz-Thompson du total est défini par

$$\hat{t}_{\text{strat}} = \sum_{h=1}^{H} \sum_{k \in s_h} \frac{Y_k}{\pi_k} = \sum_{h=1}^{H} \hat{t}_h,$$

où $\hat{t}_h$ désigne l'estimateur de Horvitz-Thompson du total de Y dans la strate U_h. L'indépendance entre les tirages permet de calculer simplement sa variance par

$$V(\hat{t}_{\text{strat}}) = \sum_{h=1}^{H} V(\hat{t}_h). \tag{1.5}$$

Par exemple, si nous effectuons un sondage aléatoire simple sans remise à l'intérieur de chaque strate nous obtenons,

$$\hat{t}_{\text{strat}} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in s_h} Y_k$$

et

$$V(\hat{t}_{\text{strat}}) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_H} \right) S_{Y,U_h}^2.$$

L'estimateur du total et sa variance dépendent alors directement de la taille n_h de l'échantillon s_h dans la strate h. Différentes méthodes ont été proposées pour choisir les tailles n_h. Une première possibilité est de considérer une allocation proportionnelle

$$\frac{n_h}{N_h} = \frac{n}{N}, \qquad h = 1, ..., H. \tag{1.6}$$

Pour utiliser cette allocation, il suffit de connaitre la taille N_h de chaque strate h ou bien N_h/N. En général, en utilisant un plan stratifié avec allocation proportionnelle on obtient de meilleurs résultats qu'avec un sondage aléatoire simple sans remise. En effet, en stratifiant, nous diminuons la variance entre les strates. Ainsi le gain est d'autant plus grand que la dispersion inter strates est grande, c'est-à-dire que les moyennes des strates sont éloignées les unes des autres.

Toutefois, l'allocation proportionnelle ne conduit généralement pas au meilleur plan de sondage stratifié. Si on souhaite estimer une moyenne ou un total, Neyman (1934) a montré qu'il existe une allocation optimale des strates pour une taille d'échantillon fixée. Il faut déterminer les tailles n_h des strates s_h telles que la variance de l'estimateur de Horvitz-Thompson soit minimale pour une taille d'échantillon et un coût d'enquête fixés. Lorsque le coût d'enquête par strate est constant, on doit donc résoudre le problème suivant

$$\min_{(n_1,\ldots,n_H)} V(\hat{t}_{\text{strat}}) \text{ sous la contrainte } \sum_{h=1}^{H} n_h = n \text{ et } n_h > 0,\ h = 1,\ldots,H$$

La solution est obtenue en introduisant un multiplicateur de Lagrange dans le problème d'optimisation (cf. par exemple Cochran (1977)) et

$$n_h = n\frac{N_h S_{Y,U_h}}{\sum_{j=1}^{H} N_j S_{Y,U_j}}, \qquad h = 1,\ldots,H. \tag{1.7}$$

La valeur des écarts-types S_{Y,U_h} n'est jamais connue. Cependant, si nous disposons d'une variable X connue sur l'ensemble de la population et très corrélée à notre variable d'intérêt, nous pouvons calculer les tailles des échantillons à sélectionner en substituant les écarts-types de la variable Y par ceux de la variable X. Ainsi la taille n_h sera définie par

$$n_h = n\frac{N_h S_{X,U_h}}{\sum_{j=1}^{H} N_j S_{X,U_j}}, \qquad h = 1,\ldots,H \tag{1.8}$$

et on parle alors d'*allocation X-optimale.*

Le gain en précision entre l'allocation optimale et l'allocation proportionnelle dépend directement de la variance des écarts-types des strates $V(S_{Y,U_h})$, $h = 1,\ldots,H$. Plus l'ampleur des dispersions change en passant d'une strate à l'autre, plus le gain dû à la stratification optimale est important (cf. Tillé (2001)).

Plan de sondage équilibré

Considérons maintenant p variables auxiliaires réelles $X_1,\ldots,X_p$ observées pour tous les individus de la population U. Soit $\mathbf{x}_k = (x_{1k},\ldots,x_{pk})'$ le vecteur des variables auxiliaires observées pour le *k-ème* individu.

Un plan de sondage $p(.)$ est dit équilibré sur les variables $X_1, ..., X_p$ si pour tout échantillon s tel que $p(s) > 0$, on a

$$\sum_{k \in s} \frac{x_{kj}}{\pi_k} = t_{X_j} \quad \forall i = 1, ..., p, \tag{1.9}$$

où t_{X_j} est le total de la variable X_j sur la population U.

Lorsqu'on effectue un tirage à probabilités égales, l'équilibrage assure la représentativité du sondage selon les variables $X_1, ..., X_p$.

Pour obtenir un échantillon équilibré, nous pouvons utiliser l'algorithme du cube développé par Deville et Tillé (2004). Si nous équilibrons uniquement sur le vecteur des probabilités d'inclusion $\boldsymbol{\pi} = (\pi_1, ..., \pi_k)$, nous obtenons un échantillon à probabilités inégales sans remise de taille fixe, puisque $\sum_{k \in s} \frac{\pi_k}{\pi_k} = n$ et $\sum_{k \in U} \pi_k = n$ (cf. Résultat 1.2).

Lorsque la taille de l'échantillon est "grande", nous pouvons approximer la variance de $\hat{t}_Y$ par

$$V(\hat{t}_Y) \simeq \sum_{k=1}^{N} \frac{1 - \pi_k}{\pi_k} \left(Y_k - \mathbf{x}_k' \boldsymbol{\beta}\right)^2$$

avec $\boldsymbol{\beta} = \left(\sum_{k=1}^{N} \frac{1-\pi_k}{\pi_k} \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_{k=1}^{N} \frac{1-\pi_k}{\pi_k} \mathbf{x}_k Y_k\right)$ (cf. Ardilly (2006), chapitre 2).

Ainsi, la variance de l'estimateur $\hat{t}_Y$ est faible lorsque les résidus pondérés de la régression linéaire des Y_k sur le vecteur $\mathbf{x}_k$ sont faibles.

Plan à probabilités inégales sans remise

Soit X une variable auxiliaire observée sur l'ensemble de la population U. On note x_k la valeur de X pour l'individu k. Si le plan est de taille fixe, la variance de l'estimateur de Horvitz-Thompson de la moyenne vaut (cf. Théorème 1.3)

$$V_p(\hat{\mu}_Y) = -\frac{1}{N^2} \frac{1}{2} \sum_{k \in U} \sum_{l \neq k \in U} \Delta_{kl} \left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2.$$

On voit que si les probabilités d'inclusion π_k sont proportionnelles aux valeurs x_k et que les x_k sont approximativement proportionnelles aux Y_k, alors Y_k/π_k sera "presque" constant et par conséquent, la variance sera très proche de zéro. L'usage des plans à probabilités inégales est particulièrement intéressant lorsque les variables sont liées par un effet de taille. Par exemple, dans notre application, nous voulons estimer la consommation moyenne sur notre population et nous disposons pour chaque individu de sa consommation antérieure. Cette variable est très corrélée à notre variable d'intérêt (cf. Figure 1.1), il sera donc intéressant de faire un échantillonnage à probabilités proportionnelles à cette variable.

Supposons que la variable auxiliaire X est à valeur positive et très corrélée positivement à la variable d'intérêt. Il est alors possible de définir les probabilités d'inclusion

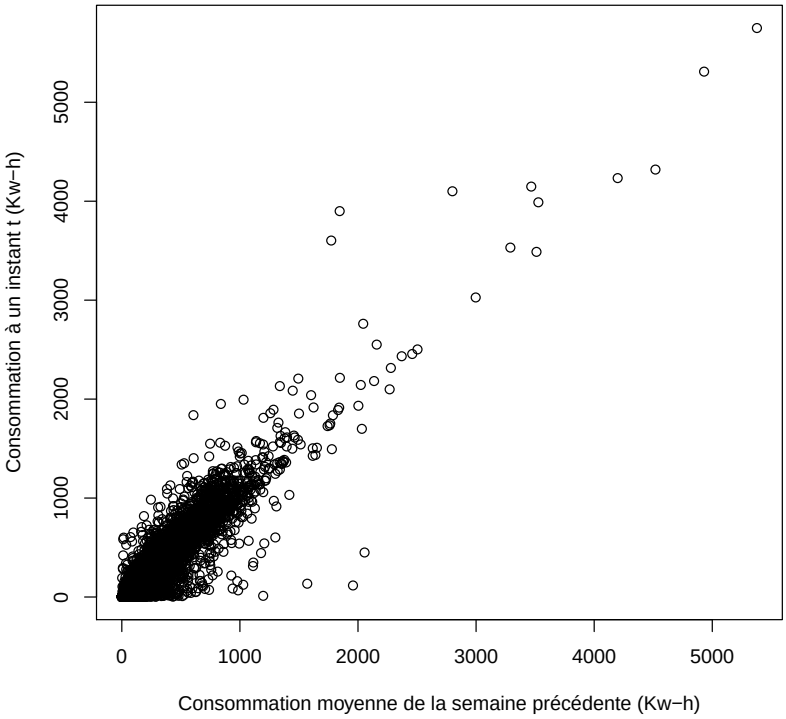


FIGURE 1.1 – Représentation de la consommation à un instant t en fonction de la consommation moyenne de la semaine précédente.

π_k par

$$\pi_k = n\frac{x_k}{\sum_{k\in U} x_k}, \quad k \in U. \tag{1.10}$$

Les plans de sondage qui vérifient cette propriété seront appelés *plans de sondage πps.*

Notons que si certaines valeurs x_k sont très élevées et n est grand, cette méthode peut conduire à des $\pi_k > 1$. Dans ce cas, nous sélectionnons automatiquement ces unités et nous recalculons les probabilités d'inclusion π_k sans les individus déjà sélectionnés. Nous répétons cet algorithme jusqu'à ce que toutes les valeurs de π_k soient inférieures ou égales à 1 (cf. Tillé (2001), chapitre 5). Nous utilisons l'estimateur de Horvitz-Thompson avec π_k définie par (1.10) pour estimer la moyenne μ_Y.

Remarque 1.2. *Si la variable auxiliaire X n'est pas liée à la variable d'intérêt ou si elle est corrélée négativement, un échantillonnage à probabilités inégales avec l'estimateur de Horvitz-Thompson peut s'avérer catastrophique (cf. Tillé (2001), chapitre 5).*

Il existe, dans la littérature, de très nombreux algorithmes d'échantillonnage à probabilités inégales qui ont été conçus pour vérifier le plus possible les 6 critères suivants (Särndal *et al.* (1992), Brewer et Hanif (1983)) :

a. La sélection de l'échantillon doit être relativement simple.
b. La taille de l'échantillon est fixée.
c. Les probabilités d'inclusion d'ordre un sont strictement proportionnelles à x_k, $k = 1, ..., N$.
d. Les probabilités d'inclusion d'ordre deux sont strictement positives (condition nécessaire pour avoir un estimateur de la variance non-biaisé).
e. Les probabilités d'ordre deux doivent être facilement calculables.
f. Les probabilités d'inclusion d'ordre deux doivent vérifier les conditions de Sen-Yates-Grundy : $\pi_{kl} - \pi_k \pi_l \leq 0 \ \forall k \neq l \in U$.

Beaucoup de méthodes ont déjà été proposées (Tillé (2006), Madow (1949)) mais aucune ne possède toutes les propriétés énoncées. Dans cette partie, nous avons décidé de détailler plus particulièrement les plans à forte entropie.

Plan de Poisson

Le plan de Poisson est la généralisation aux probabilités inégales du tirage de Bernoulli. L'échantillon est composé de tous les éléments k de U qui satisfont $\epsilon_k < \pi_k$, où, pour tout $k \in U$, les ϵ_k sont des réalisations indépendantes d'une variable aléatoire de distribution uniforme sur l'intervalle [0,1] et les π_k sont les probabilités d'inclusion du premier degré. La taille n_s de l'échantillon est aléatoire.

Les unités sont sélectionnées indépendamment les unes des autres, on a donc $\pi_{kl} = \pi_k \pi_l$ et $\Delta_{kl} = 0$ pour tous $k \neq l \in U$. Le plan de sondage est donné par

$$p(s) = \prod_{k \in s} \pi_k \prod_{k \in U - s} (1 - \pi_k).$$

L'intérêt du plan de Poisson découle de son extrême simplicité de mise en œuvre, mais aussi du fait qu'il maximise le critère d'entropie pour des probabilités d'inclusion données.

Définition 1.1. *L'entropie d'un plan est définie par*

$$I(p) = -\sum_{s \subset \mathcal{S}} p(s) \log(p(s)), \tag{1.11}$$

où on suppose que $0 \log(0) = 0$

L'entropie permet de mesurer l'indépendance entre les tirages des individus, plus elle sera élevée plus les π_{kl} seront proches de $\pi_k \pi_l$, pour $k \neq l$. Le plan de Poisson est donc le plan le plus aléatoire possible (au sens de l'entropie) qui respecte des probabilités d'inclusion d'ordre un fixées à priori (cf. Hájek (1981), chapitre 3). L'inconvénient de ce plan est qu'il ne respecte pas la condition b. La taille de l'échantillon est aléatoire et il y a une probabilité non nulle de sélectionner un échantillon vide ou l'ensemble de la population.

Tirage réjectif

Chen *et al.* (1994) et Hájek (1964) ont montré que, si les probabilités d'inclusion $\boldsymbol{\pi} = (\pi_1, .., \pi_N)$ et la taille de l'échantillon sont fixées, le tirage réjectif, noté $R(s)$, est le

plan qui a la plus forte entropie. Ce plan peut être vu comme un tirage de Poisson de probabilité d'inclusion $\boldsymbol{p} = (p_1, .., p_N)$ conditionné par la taille de l'échantillon $n_s = n$ avec $n = \sum_{k \in U} \pi_k$. Par conséquent le tirage réjectif peut-être défini par

$$R(s) = \begin{cases} c \prod_{k \in s} p_k \prod_{k \in U \setminus s} (1 - p_k) & \text{si l'échantillon est de taille } n \\ 0 & \text{sinon} \end{cases}$$

où la constante c est choisie de telle sorte que $\sum_{s \in \mathcal{S}} R(s) = 1$ avec $\mathcal{S}$ l'ensemble de tous les échantillons possibles.

Différents algorithmes existent dans la littérature pour exprimer le vecteur $\mathbf{p}$ en fonction des $\boldsymbol{\pi}$ et vice-versa (cf. Tillé (2006), Chen *et al.* (1994), Deville (2000)).

Sondage de Rao-Sampford

La première unité est sélectionnée avec des probabilités de tirage $p_k = \pi_k / n$. Les $n-1$ unités restantes sont tirées avec remise à l'aide de probabilités proportionnelles à $\pi_k / (\pi_k - 1)$.

Si toutes les unités sélectionnées sont différentes nous acceptons l'échantillon, sinon nous le rejetons et nous recommençons l'algorithme de tirage. Les probabilités d'inclusion d'ordre un sont égales à $\pi_1, ..., \pi_N$ (Hájek (1981)).

Pour de nombreux plans à probabilités inégales (tirage réjectif, tirage successif), l'estimation directe de la variance en utilisant (1.2) est impossible, car les probabilités d'inclusion d'ordre deux ne sont pas connues exactement. Särndal *et al.* (1992) proposent d'estimer la variance $V(\hat{t}_Y)$ par la variance de l'estimateur de Hansen-Hurvitz

$$V_{HH}(\hat{t}_Y) = \frac{1}{N^2} \frac{1}{n(n-1)} \sum_s \left(\frac{Y_k}{\pi_k} - \frac{1}{n} \sum_s \frac{Y_k}{\pi_k} \right)^2$$

obtenu dans le cas d'un plan avec remise (cf. Hansen et Hurvitz (1943)). L'inconvénient de cet estimateur est qu'il est généralement biaisé (cf. Särndal *et al.* (1992)).

Pour les plans à forte entropie, des formules d'approximation de la variance d'un paramètre univarié ne faisant pas intervenir les π_{kl} ont été proposées dans la littérature (Hájek (1964), Deville et Tillé (2005)). Dans le cas d'un tirage réjectif, Hájek (1964) a proposé une approximation asymptotique de la variance qui ne fait intervenir que les probabilités d'inclusion d'ordre un et qui est facile à calculer. Cette approximation est également très efficace quand le plan est proche de l'entropie maximale (cf. Berger (1998a)).

Une formule générale de l'approximation de la variance, dans le cas d'un sondage sans remise proche de l'entropie maximale, est donnée par (cf. Deville et Tillé (2005))

$$V_{approx}(\hat{t}_Y) = \sum_{k \in U} \frac{b_k}{\pi_k^2} (Y_k - Y_k^*)^2 \tag{1.12}$$

avec

$$Y_k^* = \pi_k \frac{\sum_{l \in U} b_l Y_l / \pi_l}{\sum_{l \in U} b_l}.$$

Beaucoup de variantes du paramètre b_k ont été proposées :

i. l'approximation proposée par Hájek (1964) :

$$b_k = \pi_k(1-\pi_k)\frac{N}{N-1}.$$

ii. Approximation du point fixe. Deville et Tillé (2005) ont proposé de résoudre le système suivant pour trouver une autre approximation de b_k

$$b_k - \frac{b_k^2}{\sum_{l\in U} b_l} = \pi_k(1-\pi_k).$$

A partir de la formule (1.12), nous pouvons obtenir une approximation de la variance estimée (Tillé (2001), Deville et Tillé (2005)) :

$$\hat{V}_{approx}(\hat{t}_Y) = \sum_{k\in s} \frac{c_k}{\pi_k^2}(Y_k - \hat{Y}_k^*)^2$$

avec

$$\hat{Y}_k^* = \pi_k \frac{\sum_{l\in s} c_l Y_l/\pi_l}{\sum_{l\in s} c_l}.$$

Différentes variantes pour le paramètre c_k ont été proposées :

i. Deville (1993) propose

$$c_k = (1-\pi_k)\frac{n}{n-1}.$$

ii. Berger (1998a) propose

$$c_k = (1-\pi_k)\frac{\sum_{k\in s}(1-\pi_k)}{\sum_{k\in s}\pi_k(1-\pi_k)}. \tag{1.13}$$

iii. Approximation du point fixe. Deville et Tillé (2005) ont proposé de résoudre le système suivant pour trouver une autre approximation de c_k

$$c_k - \frac{c_k^2}{\sum_{l\in s} c_l} = (1-\pi_k).$$

1.3.2 Au niveau de l'estimation

Dans cette section, nous allons présenter des méthodes qui utilisent l'information auxiliaire seulement au niveau de l'estimation. Elles permettent de redresser le poids $1/\pi_k$ des individus de l'échantillon.

La post-stratification

Supposons qu'à partir des valeurs d'une variable ou d'un croisement de variables auxiliaires qualitatives, nous pouvons déterminer H catégories, $U_1, ..., U_H$, appelées post-strates, de tailles $N_1, ..., N_H$ telles que

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

et

$$N_1 + ... + N_H = N.$$

Les tailles N_h, $h = 1, ..., H$, des post-strates dans la population sont supposées connues. Soit s un échantillon de taille fixée n, choisi aléatoirement dans U selon un plan de sondage $p(\cdot)$. Pour chaque individu $k \in s$, on détermine à quelle post-strate il appartient. Soit $s_h = s \cap U_h$, $h = 1, ..., H$, de taille aléatoire n_h connue seulement après avoir tiré l'échantillon s.

Dans le cas du sondage aléatoire simple, l'estimateur post-stratifié du total vaut :

$$\hat{t}_{POST} = \sum_{h=1}^{H} N_h \frac{1}{n_h} \sum_{k \in s_h} Y_k.$$

Cet estimateur est approximativement sans biais par rapport au vrai total si $n_h > 0$, $h = 1, ..., H$. La formule exacte de la variance est compliquée à obtenir du fait que la variable n_h est aléatoire. On peut toutefois obtenir une approximation de la variance. Par exemple, pour le sondage SRSWOR, nous obtenons

$$\begin{aligned} V(\hat{t}_{POST}) &\simeq \frac{N-n}{n} \sum_{h=1}^{H} N_h S^2_{YU_h} + \frac{N^2(N-n)}{n^2(N-1)} \sum_{h=1}^{H} \left(\frac{N-N_h}{N} \right) S^2_{YU_h} \\ &= V(\hat{t}_{PROP})[1 + O(n^{-1})], \end{aligned}$$

où $V(\hat{t}_{PROP})$ est la variance de l'estimateur de Horvitz-Thompson pour le sondage stratifié avec allocation proportionnelle et les post-strates coïncident avec les strates (cf. Tillé (2001), chapitre 10).

La stratification *a priori* permet d'obtenir une meilleur précision. Cependant, ce n'est pas toujours possible de la mettre en place, la poststratification est alors une stratégie d'estimation intéressante.

Comme le remarque Ardilly (2006), cette technique est bien adaptée au traitement des valeurs dites "extrêmes", c'est-à-dire particulièrement faibles ou particulièrement élevées. Plutôt que de supprimer ces individus, on préfère constituer des post-strates isolant les individus atypiques. On leur donne ainsi un poids différent mais ils restent dans notre échantillon (il n'y a pas d'élimination abusive). Il est toutefois nécessaire de connaitre la taille de notre post-strate dans la population.

Il est tout à fait possible d'effectuer une post-stratification avec des plans de sondage plus complexes à probabilités égales ou inégales (cf. Ardilly (2006)).

Le calage

Soit $\mathbf{x}_k = (x_{k1}, ..., x_{kp})'$ le vecteur contenant les valeurs des p variables auxiliaires mesurées sur le k-ème individu. Supposons que le total $t_{\mathbf{x}} = \sum_U \mathbf{x}_k$ est connu et que le couple $(\mathbf{x}_k, Y_k)$ est observé pour tout $k \in s$. Soit $\hat{t}_Y = \sum_s \frac{Y_k}{\pi_k}$ l'estimateur de Horvitz-Thompson du total t_Y.

Le calage consiste à chercher un nouvel ensemble de poids $\{\omega_{ks}\}_{k\in s}$ tels que les conditions suivantes soient vérifiées

i. ω_{ks} doit être aussi proche que possible du poids de sondage $d_k = 1/\pi_k$ au sens d'une distance choisie par le statisticien.

ii. L'équation suivante doit être satisfaite :

$$t_{\mathbf{x}} = \sum_U \mathbf{x}_k = \sum_{k\in s} \omega_{ks}\mathbf{x}_k.$$

Avec ces nouveaux poids, les estimateurs des totaux des variables auxiliaires sont égaux aux totaux sur toute la population.

Pour résoudre ce problème d'optimisation sous contrainte linéaire, Deville et Särndal (1992) ont appliqué la méthode des multiplicateurs de Lagrange.

Différents estimateurs de calage ont été obtenus à partir de différentes distances entre $d_k = 1/\pi_k$ et ω_{ks} dont la méthode linéaire, la méthode exponentielle, la méthode *logit* et la méthode linéaire tronquée (cf. Deville et Särndal (1992)).

Le total t_Y est alors estimé par

$$\hat{t}_{Y,\omega} = \sum_{k\in s} \omega_{ks} Y_k.$$

Deville et Särndal (1992) montre que la variance asymptotique de l'estimateur par calage $\hat{t}_{Y,\omega}$ est donnée par

$$AV(\hat{t}_{Y,\omega}) = \sum_U \sum_U (\pi_{kl} - \pi_k\pi_l)(d_k E_k)(d_l E_l),$$

où $E_k = Y_k - \mathbf{x}_k'\boldsymbol{\beta}$ et $\boldsymbol{\beta}$ satisfait l'équation

$$\left(\sum_U q_k \mathbf{x}_k \mathbf{x}_k'\right)\boldsymbol{\beta} = \sum_U q_k \mathbf{x}_k Y_k.$$

Les q_k sont des coefficients de pondération qui permettent de déterminer l'importance de chaque unité dans l'échantillon s.

L'estimateur par calage est d'autant plus efficace que les résidus E_k/π_k, $k = 1, ..., N$, sont petits.

Cette méthode est aussi utilisée pour réduire les erreurs de couverture et celles dues à la non-réponse (cf. Deville et Särndal (1992), Lundström et Särndal (1999)).

1.4 Présentation de l'approche modèle

Jusqu'à présent, nous nous sommes placés dans l'approche traditionnelle : le *design-based*. Dans celle-ci, le vecteur $\mathbf{Y} = (Y_1, ..., Y_N)$ est une quantité non aléatoire. Le seul aléa provient du plan de sondage p et les propriétés des estimateurs reposent sur le hasard engendré par la mise en œuvre du plan de sondage.

Nous allons maintenant adopter un autre point de vue et considérer que $\mathbf{Y} = (Y_1, ..., Y_N)$ est une réalisation du vecteur de variable aléatoire $\mathcal{Y} = (\mathcal{Y}_1, ..., \mathcal{Y}_N)'$ de distribution ξ. Une telle distribution de probabilité ξ est appelée *modèle de superpopulation*. Dans beaucoup de cas, la distribution de ξ est liée à une variable auxiliaire fixée $\mathbf{X} = (x_1, ..., x_N)'$ où les éléments sont supposés connus.

Dans le cadre du *modèle de superpopulation*, nous sommes en présence de deux types d'aléa : le premier est celui induit par le plan de sondage $p(.)$, le second est introduit par la variable $\mathcal{Y}$ dont la distribution dépend de l'information auxiliaire $\mathbf{X}$. Nous avons donc besoin d'introduire quelques nouvelles notations et définitions.

Soit $T = T(\mathcal{Y}_1, ..., \mathcal{Y}_N)$ une fonction de $\mathcal{Y}_1, ..., \mathcal{Y}_N$, on note $\mathbb{E}_\xi(T)$ l'espérance, $V_\xi(T)$ la variance de T par rapport au modèle ξ. Nous appelons *statistique* une fonction $T = T(\mathcal{D})$ où $\mathcal{D} = \{(k, \mathcal{Y}_k) : k \in S\}$, S étant une variable aléatoire à valeurs dans $\mathcal{S}$, l'ensemble de tous les échantillons possibles s.

La statistique T pour $S = s$ utilisant l'inférence sur la population moyenne $\mu_\mathcal{Y} = N^{-1} \sum_U \mathcal{Y}_k$ est appelée prédicteur et T pour $\mathcal{Y}_k = Y_k$ est appelée un estimateur pour $\mu_Y = N^{-1} \sum_U Y_k$ (Cassel *et al.* (1976)).

Définition 1.2.

i. T est dit sans biais par rapport au plan p (ou p-sans biais) pour la moyenne $\mu_\mathcal{Y}$ si et seulement si, pour un plan p donné, $\mathbb{E}_p(T(\mathbf{Y})) = \mu_Y$ pour tout $\mathbf{Y} = (Y_1, ..., Y_N)$.

ii. T est dit sans biais par rapport au modèle ξ (ou ξ-sans biais) pour la moyenne $\mu_\mathcal{Y}$ si et seulement si, pour un modèle ξ donné, $\mathbb{E}_\xi(T(\mu_\mathcal{Y}) - \mu_\mathcal{Y}) = 0$ pour tout $s \in \mathcal{S}$.

iii. T est dit sans biais par rapport au plan p et au modèle ξ (ou $p\xi$-sans biais) pour la moyenne $\mu_\mathcal{Y}$ si et seulement si, pour un plan p et un modèle ξ donnés, $\mathbb{E}_\xi \mathbb{E}_p(T(\mu_\mathcal{Y}) - \mu_\mathcal{Y}) = 0$.

Définition 1.3. *Un plan de sondage est dit non-informatif si et seulement si la sélection de l'échantillon ne dépend pas de la variable étudiée $\mathbf{Y}$ après avoir pris en compte l'information auxiliaire. Sous un plan de sondage non-informatif E_ξ et E_p sont commutatives c'est-à-dire que*

$$\mathbb{E}_\xi \mathbb{E}_p(T) = \mathbb{E}_p \mathbb{E}_\xi(T).$$

Considérons *le modèle de superpopulation* suivant :

$$\xi : \quad Y_k = m(x_k) + \epsilon_k, \tag{1.14}$$

où $m(.)$ est inconnue et les ϵ_k, $k = 1, ..., n$, sont des variables aléatoires indépendantes vérifiant $\mathbb{E}_\xi(\epsilon_k) = 0$ et $V_\xi(\epsilon_k) = \sigma_k^2$.

Dans ce contexte, deux approches sont possibles pour estimer un total ou une moyenne. La première dite *Model-based* est présentée de manière détaillée dans Valliant *et al.* (2000) : l'estimateur de t_Y est calculé à partir de la valeur observée pour les individus échantillonnés et de la valeur prédite pour les individus non échantillonnés. Dans ce cas, l'estimateur du total a la forme suivante :

$$\hat{t}_{Y,MB} = \sum_{k \in U \backslash s} \tilde{m}(x_k) + \sum_{k \in s} Y_k$$

où $\tilde{m}$ est l'estimation de la prédiction pour les valeurs non-échantillonnées de Y.

Remarque 1.3. *L'estimateur $\hat{t}_{Y,MB}$ est asymptotiquement non biaisé par rapport au modèle et efficace quand $m(x_k)$ et $v(x_k)$ sont correctement spécifiés. Cependant, il peut être biaisé et même non convergent si le modèle est faux (Valliant* et al. *(2000)).*

Dans la suite de ce document, nous n'allons pas développer le *Model-based approach.* Pour plus de détails, on pourra se référer à Royall (1970), Cassel *et al.* (1976) et Valliant *et al.* (2000).

La deuxième approche, dite *Model-assisted* (cf. Särndal *et al.* (1992)), s'appuie également sur la prédiction $m(x_k)$ sur toute la population. La différence provient du fait qu'elle corrige le biais possible dû à la prédiction (Breidt et Opsomer (2009), Särndal *et al.* (1992)). Ainsi l'estimateur de t_Y est de la forme :

$$\hat{t}_{Y,MA} = \sum_{k \in U} \hat{m}(x_k) + \sum_{k \in s} \frac{Y_k - \hat{m}(x_k)}{\pi_k}, \tag{1.15}$$

où $\hat{m}$ est l'estimateur de la prédiction obtenue à partir du modèle ξ et de l'échantillon s.

Considérons que nous disposons de p variables auxiliaires $\mathcal{X}_1, ..., \mathcal{X}_p$ et que les Y_k, $k = 1, ..., N$ sont indépendants. Soit $\mathbf{x}_k = (x_{1k}, ..., x_{pk})'$ le vecteur des variables auxiliaires observées pour le *k-ème* individu.

L'objectif est d'estimer le total inconnu de Y, $t_Y = \sum_U Y_k$, à partir des $(Y_k, \mathbf{x}_k)$ observés pour $k \in s$ et des $\mathbf{x}_k$ connus pour $k \in U \backslash s$ ou des totaux des variables auxiliaires $t_{\mathbf{x}} = \sum_U \mathbf{x}_k$.

Lorsque la fonction $m(.)$ est connue, Cassel *et al.* (1976) introduisent l'estimateur par la différence généralisée :

$$\begin{aligned} \hat{t}_{Y,diff} &= \sum_U m(\mathbf{x}_k) + \sum_s \frac{Y_k - m(\mathbf{x}_k)}{\pi_k} \\ &= \sum_U m(\mathbf{x}_k) + \sum_s \check{D}_k. \end{aligned} \tag{1.16}$$

Résultat 1.3. *Cassel* et al. *(1976)*

i. L'estimateur $\hat{t}_{Y,diff}$ est sans bais pour t_Y par rapport au plan p.

ii. La variance de l'estimateur $\hat{t}_{Y,diff}$ sous le plan p est égale à

$$V_p(\hat{t}_{Y,diff}) = \sum_U \sum_U \Delta_{kl} \check{D}_k \check{D}_l. \tag{1.17}$$

iii. L'estimateur non biaisé de la variance $V_p(\hat{t}_{Y,diff})$ est égal à

$$\hat{V}_p(\hat{t}_{Y,diff}) = \sum_s \sum_s \check{\Delta}_{kl} \check{D}_k \check{D}_l. \tag{1.18}$$

Dans la pratique, la fonction $m(.)$ n'est jamais connue. Des modèles paramétriques et non-paramétriques ont été proposés dans la littérature pour l'estimer.

Modèle linéaire (Särndal *et al.* (1992))

Särndal *et al.* (1992) considèrent le modèle linéaire de superpopulation ξ suivant :

$$\xi : \begin{cases} \mathbb{E}_\xi(Y_k) = \mathbf{x}_k' \boldsymbol{\beta} = \sum_{j=1}^J \beta_j x_{jk} \\ V_\xi(Y_k) = \sigma_k^2 \end{cases} \tag{1.19}$$

où $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ est inconnus et les σ_k^2 sont connus à une constante près.

Si la matrice $\sum_U \mathbf{x}_k \mathbf{x}_k' / \sigma_k^2$ est inversible, il est possible, sous le modèle ξ, d'estimer, à l'aide des moindres carrés ordinaires, le vecteur $\boldsymbol{\beta}$ par $\tilde{\boldsymbol{\beta}} = \left(\sum_U \mathbf{x}_k \mathbf{x}_k' / \sigma_k^2\right)^{-1} \sum_{k \in U} \mathbf{x}_k Y_k / \sigma_k^2$.

Les Y_k ne sont connus que pour $k \in s$. Pour obtenir une estimation de $\tilde{\boldsymbol{\beta}}$ sous le plan de sondage p, nous remplaçons chaque somme dans $\tilde{\boldsymbol{\beta}}$ par son estimateur de Horvitz-Thompson. Ainsi, si la matrice $\sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_k \sigma_k^2}$ est inversible, $\tilde{\boldsymbol{\beta}}$ est estimée par

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_k \sigma_k^2} \right)^{-1} \sum_{k \in s} \frac{\mathbf{x}_k Y_k}{\pi_k \sigma_k^2}.$$

Ainsi, pour tout $k \in U$, la prédiction $\hat{m}(\mathbf{x}_k)$ dans l'équation (1.15) devient dans le cas du modèle linéaire

$$\hat{m}(\mathbf{x}_k) = \mathbf{x}_k' \widehat{\boldsymbol{\beta}}$$

et le total t_y est estimé par

$$\widehat{t}_Y = \sum_{k \in U} \mathbf{x}_k' \widehat{\boldsymbol{\beta}} - \sum_{k \in s} \frac{\mathbf{x}_k' \widehat{\boldsymbol{\beta}} - Y_k}{\pi_k}. \tag{1.20}$$

Cet estimateur est appelé *estimateur par la régression généralisée* (GREG).

En supposant que la fonction $m(.)$ est de la forme $\sum_{j=1}^J \beta_j x_j$, Särndal *et al.* (1992) se sont placés dans un contexte paramétrique. D'une manière plus générale, il est aussi possible de travailler avec des modèles non-paramétriques, semi-paramétriques et additifs (Breidt et Opsomer (2009)).

Le modèle de superpopulation ξ dans le cadre non-paramétrique est de la forme :

$$\xi : \quad y_k = m(x_k) + \epsilon_k \tag{1.21}$$

où les ϵ_k sont des variables aléatoires indépendantes identiquement distribuées d'espérance nulle et de variance $\sigma^2 v_k$, X est une variable univariée et $m(x_k)$ est une fonction inconnue sans forme paramétrée définie a priori qui possède certaines régularités (continuité, dérivabilité, etc.). On pose $\mathbf{Y}_U = [Y_k]_{k\epsilon U}$ et $\mathbf{Y}_s = [Y_k]_{k\epsilon s}$.

Estimation non-paramétrique à l'aide des polynômes locaux (Breidt et Opsomer (2000))

Breidt et Opsomer (2000) proposent d'estimer la fonction $m(.)$ à l'aide de polynômes locaux. Cette approche consiste à effectuer en chaque point une régression linéaire pondérée, la pondération étant contrôlée par un noyau et une fenêtre associée.

Soient K une fonction noyau continue, h la fenêtre et $\mathbf{x} = (x_1, ..., x_N)'$ une variable auxiliaire.

Posons

$$\mathbf{X}_{Uk} = [1, (x_l - x_k), ..., (x_l - x_k)^q]_{l\epsilon U}\,, \qquad \mathbf{X}_{sk} = [1, (x_l - x_k), ..., (x_l - x_k)^q]_{l\epsilon s}\,,$$

$$\mathbf{e}_1 = (1, 0, ..., 0)', \qquad \mathbf{W}_{Uk} = diag\left[\frac{1}{h}K\left(\frac{x_l - x_k}{h}\right)\right]_{l\epsilon U}, \qquad \mathbf{W}_{sk\pi} = diag\left[\frac{1}{h\pi_l}K\left(\frac{x_l - x_k}{h}\right)\right]_{l\epsilon s}.$$

Ils proposent d'approximer la fonction $m(x_k)$ dans le modèle de superpopulation (1.21) par

$$m_k = m(x_k) \approx \mathbf{e}_1'(\mathbf{X}_{Uk}'\mathbf{W}_{Uk}\mathbf{X}_{Uk})^{-1}\mathbf{X}_{Uk}'\mathbf{W}_{Uk}\mathbf{Y}_U.$$

Les Y_k étant connus que pour $k \in s$, une estimation de m_k est donnée par

$$\hat{m}_k = \hat{m}(x_k) = \mathbf{e}_1'(\mathbf{X}_{sk}'\mathbf{W}_{sk\pi}\mathbf{X}_{sk})^{-1}\mathbf{X}_{sk}'\mathbf{W}_{sk\pi}\mathbf{Y}_s, \quad k \in U.$$

Estimation non-paramétrique à l'aide des splines pénalisées (Breidt *et al.* (2005))

Breidt *et al.* (2005) proposent d'estimer la fonction $m(.)$ à l'aide d'une décomposition de la fonction de régression dans une base de fonctions splines des polynômes tronqués.

Soient p le degré de la spline et $\kappa_1, ..., \kappa_L$ un ensemble fixé de nœuds. Posons :

$$\boldsymbol{\beta} = (\beta_0, ..., \beta_{p+L})', \qquad \mathbf{x}_k' = (1, x_k, ..., x_k^p, (x_k - \kappa_1)_+^p, ..., (x_k - \kappa_L)_+^p),$$

$$\mathbf{X}_s = [\mathbf{x}_k']_{k\epsilon s}, \qquad \mathbf{X}_U = [\mathbf{x}_k']_{k\epsilon U}, \qquad \mathbf{W}_s = diag(\pi_k)_{k\epsilon s}, \qquad A_\lambda = diag(0, ..., 0, \lambda, ..., \lambda).$$

Ils proposent d'approximer la fonction $m(.)$ dans le modèle de superpopulation (1.21) par

$$m(x) \approx \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{l=1}^{L} \beta_{p+l}\{(x - \kappa_l)_+\}^p, \tag{1.22}$$

avec $\boldsymbol{\beta}$ inconnu.

Pour déterminer l'estimateur de $\boldsymbol{\beta}$ sur le modèle ξ, ils cherchent à minimiser le critère des moindres carrés pénalisés, suivant

$$\sum_{k\epsilon U}(y_k - m(x_k))^2 + \lambda \sum_{l=1}^{L} \beta_{p+l}^2,$$

où $\lambda > 0$.

Si toute la population U était totalement observée, l'estimateur de $\boldsymbol{\beta}$, pour λ fixé, serait :

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}_U'\mathbf{X}_U + \mathbf{A}_\lambda)^{-1}\mathbf{X}_U'\mathbf{Y}_U \tag{1.23}$$

et

$$\tilde{m}_k = \mathbf{x}_k'\tilde{\boldsymbol{\beta}} = \mathbf{x}_k'(\mathbf{X}_U'\mathbf{X}_U + A_\lambda)^{-1}\mathbf{X}_U'\mathbf{Y}_U. \tag{1.24}$$

Dans la réalité, les $\tilde{m}_k$ ne peuvent pas être calculés, ils sont estimés par l'estimateur pondéré :

$$\hat{m}_k = \mathbf{x}_k'\hat{\boldsymbol{\beta}} = \mathbf{x}_k'(\mathbf{X}_s'\mathbf{W}_s^{-1}\mathbf{X}_s + A_\lambda)^{-1}\mathbf{X}_s'\mathbf{W}_s^{-1}\mathbf{Y}_s. \tag{1.25}$$

Estimation non-paramétrique à l'aide des B-splines (Goga (2005))

Goga (2005) se place dans le même contexte que Breidt *et al.* (2005) mais son approche repose sur une décomposition différente des splines : elle utilise la base des B-splines. Cette base est connue pour être plus stable numériquement que la base P-splines (Dierckx (1993)). Elle permet de construire un modèle linéaire dont le nombre de variables (les B-splines) n'est pas fixé à l'avance.

Soit $S_{K,m}$ l'espace des fonctions splines d'ordre m ($m \geq 2$) avec K nœuds intérieurs $0 = \xi_0 < \xi_1 < ... < \xi_K < \xi_{K+1} = 1$ et

$$S_{K,m} = \{u \in C^{m-2}[0,1] : u(x) \text{ est un polynôme de degré m-1 sur } (\xi_j, \xi_{j+1})\}.$$

L'espace $S_{K,m}$ est un espace linéaire de dimension $q = K + m$ dont une base est constituée des fonctions B-splines $(B_j(.))_{j=1}^q$ définies de la manière suivante :

$$B_j(x) = (\xi_j - \xi_{j-m}) \sum_{l=0}^{m} \frac{(\xi_{j-l} - x)_+^{m-1}}{\Pi_{r=0,r\neq l}^{m}(\xi_{j-l} - \xi_{j-r})}.$$

Chaque fonction B_j pour $j = 1, .., q$ a les nœuds $\xi_{j-m}, ..., \xi_j$ avec $\xi_r = \xi_{min(max(r,0),K+1)}$ pour $r = j - m, ..., j$ (Zhou *et al.* (1998)) ce qui implique qu'elle a comme support un nombre petit et fixé d'intervalles entre les nœuds. De plus, ces fonctions sont positives, de somme 1, $\forall x \in [0,1]$ $\sum_{j=1}^q B_j(x) = 1$.

Elle propose d'approximer la fonction $m(x_k)$ dans le modèle de superpopulation (1.21) par

$$m(x_k) \approx \sum_{j=1}^{q} \theta_j B_j(x_k), \tag{1.26}$$

avec $\theta = (\theta_1, .., \theta_q)'$ inconnu.

Si tous les éléments de la population étaient connus, on pourrait construire un estimateur $\tilde{m} \in S_{K,m}$ de la fonction m à l'aide du critère des moindres carrés, pour tout $k \in U$:

$$\tilde{m}(x_k) = \sum_{j=1}^{q} \tilde{\theta}_j B_j(x_k), \tag{1.27}$$

où

$$\tilde{\theta} = \left(\sum_U \mathbf{b}(x_k)\mathbf{b}(x_k)' \right)^{-1} \left(\sum_U \mathbf{b}(x_k) Y_k \right) \tag{1.28}$$

avec $\mathbf{b}(x_k) = (B_1(x_k), ..., B_q(x_k))'$.

Les Y_k ne sont connus que pour $k \in s$. Pour obtenir une estimation de $\tilde{\boldsymbol{\beta}}$ sous le plan de sondage p, elle remplace chaque somme dans $\tilde{\theta}$ par son estimateur de Horvitz-Thompson. Ainsi $\tilde{\theta}$ est estimé par

$$\hat{\theta} = \left(\sum_s \frac{\mathbf{b}(x_k)\mathbf{b}'(x_k)}{\pi_k} \right)^{-1} \left(\sum_s \frac{\mathbf{b}(x_k) Y_k}{\pi_k} \right) \tag{1.29}$$

et

$$\hat{m}_k = \mathbf{b}'(x_k)\hat{\theta}, \qquad k \in U. \tag{1.30}$$

Pour mettre en place ces méthodes non-paramétriques, il est nécessaire de connaitre l'information auxiliaire pour tous les individus de la population. Leur généralisation au cas multivarié est en principe possible, mais le nombre de variables rend cela impraticable pour plus de deux ou trois dimensions. On parle alors du "fléau de la dimension". D'autres approches ont été proposées pour prendre en compte l'ensemble des variables auxiliaires (cf. Breidt et Opsomer (2009)) en considérant des classes de modèle plus restreintes (modèle additif, modèle semi-paramétrique, modèle additif généralisé).

1.5 Asymptotique en théorie des sondages

En général dans les enquêtes par sondage, on ne se borne pas à fournir exclusivement la valeur de l'estimateur, on cherche également à mesurer sa qualité en construisant des intervalles de confiance contenant la vraie valeur avec un niveau de confiance fixé. Pour construire de tels intervalles, il faut connaitre la distribution au moins asymptotique de l'estimateur. En théorie des sondages, la taille de la population est finie. On ne peut donc pas travailler avec la même notion d'asymptotique qu'en statistique inférentielle classique. Par ailleurs, quand la taille de l'échantillon est fixée, la sélection des éléments d'un échantillon se fait de façon dépendante (par exemple on peut s'interdire de sélectionner deux fois un même individu). Il sera nécessaire d'introduire des conditions sur la dépendance via les probabilités d'inclusion pour démontrer les propriétés asymptotiques.

Les résultats asymptotiques nécessitent de travailler avec une suite croissante de population telle que à la fois n et N tendent vers l'infini.

Cadre asymptotique de superpopulation

Nous plaçons dans le cadre asymptotique de superpopulation introduit par Isaki et Fuller (1982). Considérons une suite croissante de populations emboitées $U_1 \subset ... \subset U_\nu \subset ...$ de tailles $N_1 < ... < N_\nu < ....$ Soit $\mathcal{F}_\nu = (Y_{1\nu}, ..., Y_{\nu\nu})$ le vecteur de la variable d'intérêt sur la population U_ν.

Pour décrire la suite $\{\mathcal{F}_\nu\}$, il existe deux possibilités (cf. Fuller (2009a)). Soit on considère que $\mathcal{F}_\nu$ est une suite de nombres qui vérifient certaines propriétés (moyenne qui converge vers une valeur finie, etc.). Soit on considère que les $Y_{i\nu}$, $i = 1, ..., \nu$, sont des variables aléatoires indépendantes dont la loi vérifie certaines propriétés (ex : espérance et variance finies).

Pour chaque population $U_\nu = \{1, ..., \nu\}$, on considère le plan de sondage $p_\nu(.)$ qui assigne une certaine probabilité $p_\nu(s_\nu)$ à chaque échantillon s_ν de U_ν. On désigne par θ_ν le paramètre d'intérêt que l'on souhaite estimer.

Soient $\pi_{k\nu}$ et $\pi_{kl\nu}$ les probabilités d'inclusion d'ordre un et respectivement deux. Nous supposons également que la taille n_ν de l'échantillon s_ν n'est pas aléatoire et que $n_1 < n_2 < ....$ Soient $f_\nu = n_\nu / N_\nu$ le taux de sondage et $S_{Y\nu}$ l'écart-type de la variable d'intérêt sur la population U_ν.

On note respectivement $\mu_{Y\nu}$ et $\hat{\mu}_{Y\nu}$ la moyenne de la variable Y sur la population U_ν et son estimation sur l'échantillon s_ν. Pour simplifier les notations, on omet l'indice ν lorsqu'il n'y a pas d'ambiguïté.

Convergence asymptotique

Définition 1.4. *Un estimateur $\hat{\theta}_\nu$ de θ_ν est asymptotiquement sans biais, si*

$$\lim_{\nu \to \infty} [E_{p_\nu}(\hat{\theta}_\nu) - \theta_\nu] = 0. \tag{1.31}$$

Définition 1.5. *Un estimateur $\hat{\theta}_\nu$ est dit convergent en probabilité pour θ_ν si pour tout $\epsilon > 0$*

$$\lim_{\nu \to \infty} \mathbb{P}(|\hat{\theta}_\nu - \theta_\nu| > \epsilon) = 0. \tag{1.32}$$

Sous certaines hypothèses sur les probabilités d'inclusion et sur les moments de Y, Isaki et Fuller (1982) ont montré que l'estimateur de Horvitz-Thompson du total est convergent. Robinson et Särndal (1983) ont également établit la convergence de l'estimateur GREG défini par (1.20).

Théorème central limite

La normalité asymptotique de l'estimateur de Horvitz-Thompson n'a été montrée que pour certains plans de sondage. Ce résultat a d'abord été prouvé par Erdös et Rényi (1959) et Hàjek (1960) pour le sondage aléatoire simple sans remise.

Théorème 1.4. *Hàjek (1960)*

Supposons que $n_\nu \to \infty$ et $N_\nu - n_\nu \to \infty$. Alors dans le cas d'un sondage aléatoire simple

$$\sqrt{n_\nu}\frac{\hat{\mu}_{Y\nu} - \mu_{Y\nu}}{\sqrt{1-f_\nu}S_{Y\nu}} \to \mathcal{N}(0,1) \text{ quand } \nu \to \infty$$

si et seulement si la suite $(Y_{k\nu})$ vérifie la condition de Lindeberg-Hájek

$$\lim_{\nu\to\infty} \sum_{T_\nu(\delta)} \frac{Y_{k\nu} - \mu_{Y\nu}}{(N_\nu - 1)S_{Y\nu}^2} = 0 \text{ quel que soit } \delta > 0,$$

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

$$|Y_{k\nu} - \mu_{Y\nu}|/\sqrt{1-f_\nu}S_{Y\nu} > \delta\sqrt{n_\nu}.$$

La normalité asymptotique de l'estimateur de Horvitz-Thompson a également été étudiée pour certains plans à probabilités inégales. Par exemple, Hájek (1964) l'a démontré pour le tirage réjectif, Víšek (1979) pour le tirage de Sampford et Bickel et Freedman (1984), Krewski et Rao (1981) pour le sondage aléatoire simple stratifié et Berger (1998b) pour les plans à forte entropie.

Intervalle de confiance

Un intervalle de confiance de niveau asymptotique $1-\alpha$ pour la moyenne μ_Y est donné par

$$IC(1-\alpha) = \left[\hat{\mu}_Y - z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\mu}_Y)}, \hat{\mu}_Y + z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\mu}_Y)}\right]$$

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

1.6 Calcul de la précision

La formule d'estimation de la variance présentée dans la section 1.2 ne peut s'appliquer qu'à des fonctionnelles linéaires. On peut être dans l'incapacité de donner une expression exacte de la variance lorsqu'on s'intéresse à des paramètres d'intérêt plus complexes (ex : le ratio, le coefficient de corrélation) ou lorsqu'on est en présence de non-réponse. Toutefois, différentes techniques basées sur la linéarisation ou sur le rééchantillonnage ont été proposées pour calculer la précision d'un estimateur complexe (Särndal *et al.* (1992), Shao et Tu (1995), Deville (1999), Rao et Wu (1985), Demnati et Rao (2004)).

1.6.1 La technique de linéarisation

Cette méthode utilise le développement de Taylor pour obtenir une estimation de la variance d'une fonction f de totaux $\theta = f(t_{Y_1}, ..., t_{Y_p})$ supposée dérivable (par exemple un ratio). Un estimateur simple $\hat{\theta}$ de θ est obtenu en substituant les totaux sur la population par leurs estimateurs de Horvitz-Thompson :

$$\hat{\theta} = f(\hat{t}_{Y_1}, ..., \hat{t}_{Y_p}).$$

La méthode consiste à approximer l'estimateur non-linéaire $\hat{\theta}$ par un estimateur linéaire $\hat{\theta}_0$ construit à partir du développement de Taylor de la fonction f au point $(t_{Y_1}, ..., t_{Y_p})$. Ainsi, on peut écrire

$$\hat{\theta} \simeq \hat{\theta}_0 = \theta + \sum_{j=1}^{p} a_j(\hat{t}_{Y_j} - t_{Y_j}),$$

où $a_j = \frac{\partial f}{\partial \hat{t}_{Y_j}}\Big|_{(\hat{t}_{Y_1},...,\hat{t}_{Y_p})=(t_{Y_1},...,t_{Y_p})}$. La variance de $\hat{\theta}$ est alors approximée par celle de $\hat{\theta}_0$ (cf. Särndal *et al.* (1992))

$$V_{approx}(\hat{\theta}) = \sum_U \sum_U (\pi_{kl} - \pi_k \pi_l) \frac{u_k}{\pi_k} \frac{u_l}{\pi_l}$$

avec $u_k = \sum_{j=1}^{p} a_j Y_{jk}$ et un estimateur convergent de la variance est donné par

$$\hat{V}_{approx}(\hat{\theta}) = \sum_s \sum_s \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l}$$

où $\hat{u}_k = \sum_{j=1}^{p} \hat{a}_j Y_{jk}$ et $\hat{a}_j$ est obtenu en remplaçant chaque total de a_j par l'estimateur de Horvitz-Thompson correspondant.

1.6.2 Le Jackknife

Cette méthode a été introduite en statistique classique par Quenouille (1949) pour l'estimation du biais d'une statistique puis reprise par Tuckey (1958) pour estimer la précision d'un estimateur θ. La méthode consiste à supprimer à tour de rôle chaque individu de l'échantillon et à recalculer la statistique d'intérêt sur les individus restants. La dispersion des statistiques Jackknife est alors utilisée comme estimateur de la variance (Shao et Tu (1995)). Plus précisément, soit $\hat{\theta}$ l'estimateur de θ obtenu à partir de l'échantillon s. Pour $j = 1, ..., n$, on note $\hat{\theta}(j)$ l'estimateur obtenu en supprimant l'individu j. L'estimateur Jackknife de θ est égal à

$$\bar{\hat{\theta}}^* = \frac{1}{n} \sum_{j=1}^{n} \hat{\theta}_j^*$$

avec $\hat{\theta}_j^* = n\hat{\theta} - (n-1)\hat{\theta}(j)$. La variance inconnue $V(\hat{\theta})$ est alors estimée par

$$\hat{V}_J(\hat{\theta}) = \frac{1}{n(n-1)} \sum_{j=1}^{n} (\hat{\theta}_j^* - \bar{\hat{\theta}}^*)^2.$$

Une présentation plus détaillée de cette méthode est donnée dans Shao et Tu (1995).

D'un point de vue pratique, cette méthode est très gourmande en temps de calculs (Rao et Wu (1985)) et elle nécessite une hypothèse sur la loi de $\hat{\theta}$ pour construire un intervalle de confiance, en pratique c'est une loi normale (cf. Ardilly (2006), chapitre 5). L'utilisation du Jackknife est complexe dans le cadre de population finie. Dans le cas d'un sondage aléatoire simple sans remise, l'estimateur $\bar{\hat{\theta}}^*$ est égal à $\hat{\theta}$ mais la variance Jaccknife doit être ajustée car elle ne prend pas en compte la correction de population finie (cf. (Ardilly, 2006)). Berger et Rao (2006) ont proposé un estimateur Jackknife dans le cas d'un plan à probabilités inégales. Pour estimer la variance de quantiles, nous pouvons utiliser le delete d-jackknife (Shao et Tu (1995)).

1.6.3 Le bootstrap

C'est sans doute la méthode par réplication la plus générale. Elle a été introduite par Efron (1979) dans le cadre d'une population infinie puis adaptée dans un cadre de population finie pour des plans plus ou moins complexes (cf. Shao et Tu (1995), Davison et Hinkley (1997), Chauvet (2007)). Les méthodes de bootstrap reposent sur le principe du plug-in : la distribution de la statistique d'intérêt est inconnue, et, plutôt que de l'estimer grâce à une loi paramétrique, on va l'estimer par la loi empirique constatée sur les données. En pratique, on tire un grand nombre B de rééchantillons s^* à partir de l'échantillon s de départ et on calcule l'estimateur $\hat{\theta}^*$ du paramètre d'intérêt θ pour chacun de ces échantillons. Ces estimateurs bootstrapés permettent d'obtenir une approximation de la loi de l'estimateur du paramètre d'intérêt. On peut alors estimer la variance par

$$\hat{V}(\hat{\theta}) = \frac{1}{B-1}\sum_{i=i}^{B}(\hat{\theta}_i^* - \hat{\theta}_{(.)}^*)^2, \tag{1.33}$$

où $\hat{\theta}_{(.)}^* = \frac{1}{B}\sum_{i=i}^{B}\hat{\theta}_i^*$

Dans la suite de cette section, nous allons présenter trois des principales techniques de bootstrap existantes dans le cas d'un sondage aléatoire simple sans remise. La plupart peuvent être généralisées à des plans de sondage plus complexes (Chauvet (2007)). On désigne par s l'échantillon de taille n tiré par sondage aléatoire simple sans remise dans la population U. Soit $\hat{\mu}_Y$ l'estimateur de Horvitz-Thompson de la moyenne μ_Y et $f = n/N$ le taux de couverture.

Rescaled boostrap (Rao et Wu (1988))

Le *rescaled Bootstrap* consiste à tirer un échantillon s^* de taille m dans s par sondage aléatoire simple avec remise. L'estimateur bootstrapé de la moyenne est alors défini par

$$\hat{\mu}_Y^* = \frac{1}{m}\sum_{s^*}\tilde{Y}_k \tag{1.34}$$

avec

$$\tilde{Y}_k = \mu_Y + \left[\frac{m(1-f)}{n-1}\right]^2 (Y_k - \mu_Y), \quad k \in s^*.$$

Les valeurs réajustées $\tilde{Y}_k$ permettent de fournir une estimation sans biais de la variance. Afin d'obtenir un estimateur sans biais du moment d'ordre 3, Rao et Wu (1988) suggèrent d'utiliser

$$m = \frac{1-f}{(1-2f)^2}\frac{(n-2)^2}{n-1}.$$

L'inconvénient de cette méthode est que le réajustement peut conduire à des valeurs incohérentes pour la statistique bootstrapée (cf. Sitter (1992)).

Le Mirror-Match Bootstrap (Sitter (1992))

L'idée de cet algorithme est de constituer le ré-échantillon s^* à partir de tirage répéter dans l'échantillon s. Soit n' un entier tel que $1 \leq n' < n$. On pose $f^* = n'/n$ et $k' = \frac{n(1-f^*)}{n'(1-f)}$. On suppose que k' est entier. On sélectionne k' échantillons $s_1^*, ..., s_{k'}^*$ de taille n' dans s par sondage aléatoire simple sans remise. Le ré-échantillon est obtenu en réunissant $s_1^*, ..., s_{k'}^*$ et l'estimateur bootstrapé de la moyenne est égal à

$$\hat{\mu}_Y^* = \frac{1}{k'n'} \sum_{s^*} Y_k. \tag{1.35}$$

Sitter (1992) suggère d'utiliser $n' = fn$, c'est-à-dire d'échantillonner avec le même taux de sondage qu'au départ. Quand ce nombre est entier, ce choix permet d'obtenir un rééchantillon s^* de même taille que s.

Le bootstrap sans remise (Gross (1980))

Cette méthode consiste à générer une pseudo population U^* en répliquant chaque élément de l'échantillon s l fois où $l = 1/f$. On sélectionne alors un ré-échantillon s^* de taille n dans U^* par sondage aléatoire simple sans remise et l'estimateur bootstrapé de la moyenne est égale à

$$\hat{\mu}_Y^* = \frac{1}{n} \sum_{s^*} Y_k. \tag{1.36}$$

Différentes variantes de cette méthode ont été proposées pour tenir du compte du fait que N/n n'est pas un entier (cf. Booth *et al.* (1994), Bickel et Freedman (1984)).

1.7 Sondage sur données fonctionnelles

Nous allons maintenant nous intéresser à des observations Y_k qui sont des courbes (cf. Figure 1.2) et présenter l'estimateur de Horvitz-Thompson fonctionnel. Nous définissons ensuite les notions de convergence dans l'espace fonctionnel.

1.7.1 Estimateur de Horvitz-Thompson pour données fonctionnelles

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

$$\mu(t) = \frac{1}{N} \sum_{k\in U} Y_k(t). \tag{1.37}$$

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

$$\hat{\mu}(t) = \frac{1}{N} \sum_{k\in s} \frac{Y_k(t)}{\pi_k} = \frac{1}{N} \sum_{k\in U} \frac{Y_k(t)}{\pi_k} \mathbb{1}_{k\in s}, \quad t \in [0, T], \tag{1.38}$$

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

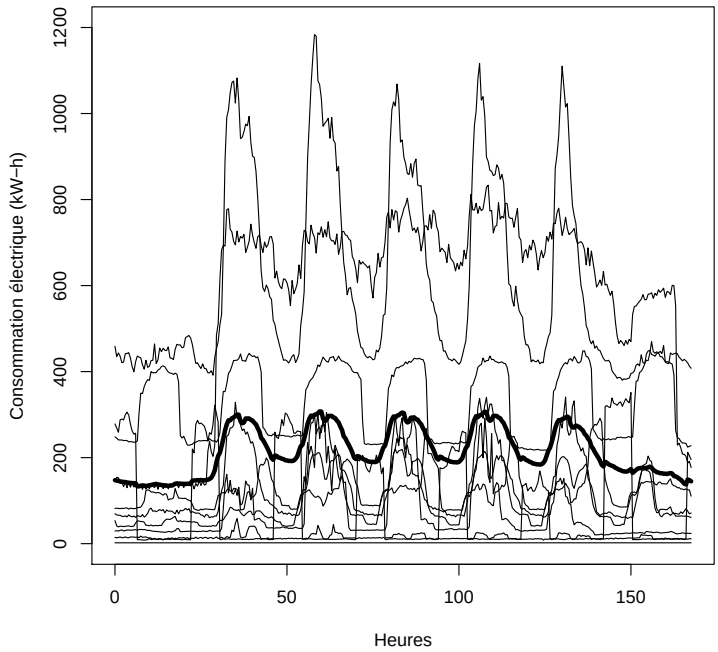


FIGURE 1.2 – Echantillon de 10 courbes de consommation électrique. La courbe moyenne est tracée en gras.

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

$$\gamma_p(r,t) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{Y_k(r)}{\pi_k} \frac{Y_l(t)}{\pi_l} \tag{1.39}$$

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

$$\widehat{\gamma}_p(r,t) = \frac{1}{N^2} \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{Y_k(r)}{\pi_k} \frac{Y_l(t)}{\pi_l} \tag{1.40}$$

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

Exemple : Sondage aléatoire simple sans remise

On considère un sondage aléatoire simple sans remise de taille n dans la population U de taille N. Dans ce cas, l'estimateur de Horvitz-Thompson pour la courbe moyenne μ défini dans (1.38) devient

$$\widehat{\mu}_{\mathrm{srswor}}(t) = \frac{1}{n} \sum_{k \in s} Y_k(t), \quad t \in [0,T]. \tag{1.41}$$

L'estimateur de la fonction de covariance défini par (1.40) est alors

$$\widehat{\gamma}_{\text{srswor}}(r,t) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{Y(r)Y(t),s}, \qquad r,t \in [0,T], \tag{1.42}$$

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

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

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

$$\widehat{\gamma}_{\text{srswor}}(t) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{Y(t),s}, \tag{1.43}$$

où

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

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

Pour les échantillons de taille fixe tirés sans remise, il est possible de donner l'équivalent de la formule de Yates et Grundy (1953) et Sen (1953),

$$\gamma_p(r,t) = -\frac{1}{2}\frac{1}{N^2} \sum_{k \in U} \sum_{l \in U, l \neq k} (\pi_{kl} - \pi_k \pi_l) \left(\frac{Y_k(r)}{\pi_k} - \frac{Y_l(r)}{\pi_l}\right)\left(\frac{Y_k(t)}{\pi_k} - \frac{Y_l(t)}{\pi_l}\right), \tag{1.44}$$

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

La formule indique clairement que si, pour tout $t \in [0,T]$ et $k \in U$, la probabilité d'inclusion du premier ordre est approximativement proportionnelle à $Y_k(t)$, la variance de l'estimateur $\hat{\mu}$ sera faible.

1.7.2 Estimation à partir de trajectoires discrétisées

Généralement les trajectoires $Y_k(t)$ ne sont pas observées continûment pour $t \in [0,T]$ mais uniquement sur un ensemble de D instants de mesure $0 = t_1 < t_2 < ... < t_D = T$. Une stratégie classique en analyse des données fonctionnelles consiste à effectuer une interpolation ou un lissage des trajectoires discrétisées afin d'obtenir des objets qui sont réellement des fonctions (Ramsay et Silverman (2005)). Cela permet également de traiter des courbes dont les instants de mesure ne sont pas identiques. Dans le cadre des sondages, l'interpolation linéaire, lorsqu'il n'y a pas d'erreur de mesure aux points discrétisés, a été étudiée par Cardot et Josserand (2011) tandis que des procédures de lissage sont proposées dans Cardot *et al.* (2012). Si le nombre de points de discrétisation est suffisant et si les trajectoires sont assez régulières (mais pas nécessairement dérivables), alors l'erreur d'approximation due au lissage ou à l'interpolation est négligeable face à l'erreur d'échantillonnage.

Par la suite, nous supposerons que pour chaque unité k dans l'échantillon la courbe interpolée est définie par

$$Y_{k,d}(t) = Y_k(t_i) + \frac{Y_k(t_{i+1}) - Y_k(t_i)}{t_{i+1} - t_i}(t - t_i), \quad t \in [t_i, t_{i+1}]. \tag{1.45}$$

Il est alors possible de définir l'estimateur de Horvitz-Thompson de la courbe moyenne basé sur les trajectoires interpolées

$$\hat{\mu}_d(t) = \frac{1}{N}\sum_{k\in s}\frac{Y_{k,d}(t)}{\pi_k}, \quad t \in [0,T]. \tag{1.46}$$

La covariance γ_p sera alors estimée par

$$\widehat{\gamma}_d(r,t) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{Y_{k,d}(r)}{\pi_k}\frac{Y_{l,d}(t)}{\pi_l} \quad (r,t) \in [0,T]\times[0,T]. \tag{1.47}$$

1.7.3 Quelques rappels de probabilité

Pour construire une bande de confiance de niveau $1-\alpha$ de la forme

$$\mathbb{P}\left(\mu(t) \in \left[\widehat{\mu}(t) \pm c_\alpha \frac{\widehat{\sigma}(t)}{\sqrt{n}}\right], \ \forall t \in [0,T]\right) = 1-\alpha, \tag{1.48}$$

autour de notre courbe estimée $\hat{\mu}_d$, nous devons comme dans le cadre univarié déterminer la loi asymptotique de notre estimateur. Pour cela, nous allons considérer le cadre de superpopulation introduit par Isaki et Fuller (1982) (cf. section 1.5 pour plus de détails). Lorsque nous travaillons avec des données fonctionnelles, la convergence ponctuelle ne suffit pas. Nous devons montrer dans un premier temps que l'estimateur de la courbe moyenne est uniformément convergent puis déterminer sa convergence en distribution dans l'espace des fonctions continues sur l'intervalle $[0,T]$ muni de la norme sup, noté $C[0,T]$.

Définition 1.6. *Un estimateur $\hat{\theta}$ est dit uniformément convergent en probabilité si pour tout $\epsilon > 0$*

$$\mathbb{P}(\sup_{t\in[0,T]}|\hat{\theta}(t) - \theta(t)| > \epsilon) \to 0.$$

On peut également définir la convergence en loi dans $C[0,T]$

Définition 1.7. *Soient X_n et X des variables aléatoires à valeur dans $C[0,T]$, on dit que la suite (X_n) converge en distribution vers X dans l'espace des fonctions continues sur l'intervalle $[0,T]$ muni de norme sup, noté $C[0,T]$, si pour toute fonctionnelle $\phi : C[0,T] \to \mathbb{R}$ bornée et uniformément continue*

$$\mathbb{E}(\phi(X_n)) \to \mathbb{E}(\phi(X)), \quad \textit{quand } n \textit{ tend à l'infini.}$$

Pour démontrer la convergence en distribution, il suffira de montrer que le vecteur $(X_n(t_1), ..., X_n(t_p))$ converge en distribution vers $(X(t_1), ..., X(t_p))$ et que la suite (X_n) est tendue (Théorème 8.1 p 54 Billingsley (1968)).

Définition 1.8. *Une mesure de probabilité P sur $(C[0,T],\mathcal{C})$ est dite tendue si pour tout réel ϵ strictement positif il existe un compact K tel que*

$$P(K) > 1 - \epsilon$$

avec $\mathcal{C}$ ensemble des boréliens de $C[0,T]$.

Définition 1.9. *Si X_n sont des éléments aléatoires de $C[0,T]$, on dit que la suite $\{X_n\}$ est tendue quand $\{P_n\}$ est tendue, où P_n représente la distribution de X_n.*

Dans la pratique ces définitions ne sont pas faciles à manipuler. Pour vérifier le critère de tension nous utiliserons le théorème suivant.

Théorème 1.5. (Théorème 12.3, Billingsley (1968)). *La suite $\{X_n\}$ est tendue dans $C[0,T]$ si elle satisfait les 2 conditions suivantes :*

i. La suite $\{X_n(0)\}$ est tendue.

ii. Il existe deux constantes $\zeta \geq 0$ et $\alpha > 1$ et une fonction F continue et non décroissante sur $[0,T]$ telles que

$$\mathbb{P}\{|X_n(t_2) - X_n(t_1)| \geq \lambda\} \leq \frac{1}{\lambda^\zeta}|F(t_2) - F(t_1)|^\alpha \tag{1.49}$$

pour tous t_1, t_2 et n et tout λ positif. La condition sur les moments

$$\mathbb{E}\{|X_n(t_2) - X_n(t_1)|^\zeta\} \leq |F(t_1) - F(t_2)|^\alpha \tag{1.50}$$

implique (1.49).

Pour montrer que la suite est tendue, on s'efforcera de vérifier l'inégalité (1.50) en prenant $F(t) = t$.

Remarque 1.4. *Une technique simple pour montrer la convergence uniforme de X_n vers X consistera à montrer la convergence en loi de $X_n - X$ vers 0.*

Chapitre 2

Estimation d'une trajectoire moyenne à l'aide d'un modèle de régression linéaire fonctionnelle

Lorsque nous travaillons avec des données fonctionnelles, nous pouvons également utiliser l'information auxiliaire au niveau de l'estimation pour améliorer la précision de notre estimateur. Dans ce chapitre, nous proposons d'estimer la courbe moyenne à l'aide d'un estimateur basé sur un modèle de régression fonctionnelle (Faraway (1997), Ramsay et Silverman (2005)). Celui-ci peut être vu comme une extension directe au cadre fonctionnel de l'estimateur GREG étudié par Robinson et Särndal (1983) et Särndal *et al.* (1992). L'objectif de ce chapitre est d'établir les propriétés asymptotiques de l'estimateur de la courbe moyenne assisté par ce type de modèle.

Dans la section 2.1, nous allons introduire les notations et suggérer une légère modification de l'estimateur qui permettra de contrôler la variance des estimateurs des coefficients de régression. Sous de faibles hypothèses sur le plan de sondage et la régularité des trajectoires, nous montrons, dans la section 2.2, la convergence uniforme de notre estimateur ainsi que celle de sa fonction de variance estimée. Enfin, sous des hypothèses supplémentaires, nous démontrons, dans la section 2.3, le théorème central limite fonctionnel. Les preuves sont regroupées dans la section 2.4.

2.1 Estimation de la courbe moyenne à l'aide d'un modèle linéaire fonctionnel

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

$$\mu(t) = \frac{1}{N} \sum_{k\in U} Y_k(t). \tag{2.1}$$

Soit s un échantillon de taille fixée n, choisi aléatoirement dans U selon un plan de sondage $p(\cdot)$. Nous supposons que les probabilités d'inclusion du premier et du second

ordre satisfont $\pi_k = \mathbb{P}(k \in s) > 0$, pour tout $k \in U$, et $\pi_{kl} = \mathbb{P}(k\&l \in s) > 0$ pour tout $k, l \in U$, $k \neq l$.

Considérons p variables auxiliaires réelles $X_1, ..., X_p$ observées pour tous les individus $k \in U$ et soit x_{kj} la valeur de la variable X_j pour le k-ème individu. Notons par $\mathbf{x}_k = (x_{k1}, ..., x_{kp})'$ le vecteur contenant les valeurs des p variables auxiliaires mesurées sur le k-ème individu. Afin d'améliorer la précision de l'estimateur de la courbe moyenne μ, nous allons introduire un modèle de régression basé sur ces variables auxiliaires. Par analogie au cadre non-fonctionnel (Särndal *et al.* (1992)), nous supposons que la relation entre la variable d'intérêt fonctionnelle et les variables auxiliaires est modélisée par un modèle de superpopulation ξ défini comme suit :

$$\xi: \quad Y_k(t) = \mathbf{x}_k'\boldsymbol{\beta}(t) + \epsilon_{kt}, \quad t \in [0, T], \tag{2.2}$$

où $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))'$ est le vecteur des coefficients de régression fonctionnels, $\mathbb{E}_\xi(\epsilon_k) = 0$, les ϵ_{kt} sont indépendants pour $k \neq l$ et de fonction de covariance $\mathrm{Cov}_\xi(\epsilon_{kt}, \epsilon_{kr}) = \Gamma(t, r)$ si $k = l$ et 0 sinon, pour $(t, r) \in [0, T] \times [0, T]$. Ce modèle est une extension directe à plusieurs variables du modèle linéaire fonctionnel proposé par Faraway (1997).

Si $\mathbf{x}_k$ et Y_k sont connus pour tous les individus $k \in U$ et si la matrice $\mathbf{G} = \frac{1}{N}\sum_U \mathbf{x}_k\mathbf{x}_k'$ est inversible, il est possible d'estimer sous le modèle ξ, à l'aide des moindre carrés ordinaires, le vecteur $\boldsymbol{\beta}(t)$ par $\tilde{\boldsymbol{\beta}}(t) = \mathbf{G}^{-1}\frac{1}{N}\sum_{k\in U}\mathbf{x}_k Y_k(t)$. La courbe moyenne $\mu(t)$ peut alors être estimée par l'estimateur de la différence (cf. équation (1.16)) proposé par Cassel *et al.* (1976), pour tout $t \in [0, T]$,

$$\begin{aligned} \tilde{\mu}(t) &= \frac{1}{N}\sum_{k\in U}\mathbf{x}_k'\tilde{\boldsymbol{\beta}}(t) - \frac{1}{N}\sum_{k\in s}\frac{\mathbf{x}_k'\tilde{\boldsymbol{\beta}}(t) - Y_k(t)}{\pi_k} \\ &= \frac{1}{N}\sum_{k\in U}\tilde{Y}_k(t) - \frac{1}{N}\sum_{k\in s}\frac{\tilde{Y}_k(t) - Y_k(t)}{\pi_k}, \end{aligned} \tag{2.3}$$

où $\tilde{Y}_k(t) = \mathbf{x}_k'\tilde{\boldsymbol{\beta}}(t)$.

Dans la pratique, les couples $(\mathbf{x}_k, Y_k)$ ne sont connus que pour les individus $k \in s$. Pour obtenir un estimateur de $\mu(t)$, nous remplaçons chaque somme dans $\tilde{\boldsymbol{\beta}}(t)$ par son estimateur de Horvitz-Thompson. Ainsi, si la matrice $\widehat{\mathbf{G}} = \frac{1}{N}\sum_{k\in s}\frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k}$ est inversible, $\tilde{\boldsymbol{\beta}}(t)$ est estimé sous le plan de sondage p par

$$\widehat{\boldsymbol{\beta}}(t) = \widehat{\mathbf{G}}^{-1}\frac{1}{N}\sum_{k\in s}\frac{\mathbf{x}_k Y_k(t)}{\pi_k}, \quad t \in [0, T]. \tag{2.4}$$

Nous obtenons finalement l'estimateur de $\mu(t)$ assisté par le modèle ξ en remplaçant $\tilde{\boldsymbol{\beta}}(t)$ par $\widehat{\boldsymbol{\beta}}(t)$ dans (2.3),

$$\widehat{\mu}_{MA}(t) = \frac{1}{N}\sum_{k\in U}\widehat{Y}_k(t) - \frac{1}{N}\sum_{k\in s}\frac{\widehat{Y}_k(t) - Y_k(t)}{\pi_k}, \quad t \in [0, T], \tag{2.5}$$

où $\widehat{Y}_k(t) = \mathbf{x}_k'\widehat{\boldsymbol{\beta}}(t)$. Pour construire l'estimateur $\widehat{\mu}_{MA}(t)$, il n'est pas nécessaire d'observer le vecteur $\mathbf{x}_k$ pour l'ensemble des individus $k \in U$. Etant donné que $\sum_{k\in U} \widehat{Y}_k(t) = (\sum_{k\in U} \mathbf{x}_k)' \widehat{\boldsymbol{\beta}}(t)$, il suffit de connaitre $\mathbf{x}_k$ et $Y_k(t)$ pour l'ensemble des individus $k \in s$ ainsi que le total de chaque variable auxiliaire sur l'ensemble de la population $\sum_{k\in U} \mathbf{x}_k$.

Remarque 2.1. *Si le vecteur des variables auxiliaires contient l'intercept, alors l'estimateur de Horvitz-Thompson des résidus estimés $\widehat{Y}_k(t) - Y_k(t)$ est nul à chaque instant $t \in [0, T]$ (cf. Särndal (1980)). Dans ce cas, l'estimateur $\widehat{\mu}_{MA}$ est la moyenne sur toute la population U des valeurs estimées $\widehat{Y}_k(t)$ par le modèle*

$$\widehat{\mu}_{MA}(t) = \frac{1}{N} \sum_U \widehat{Y}_k(t), \quad t \in [0, T]. \tag{2.6}$$

De plus, si l'intercept est l'unique variable auxiliaire utilisée dans le modèle, c'est-à-dire $Y_k(t) = \beta(t) + \varepsilon_{kt}$ pour tout $k \in U$, alors l'estimateur $\hat{\mu}_{MA}$ est égal à l'estimateur de Hájek,

$$\widehat{\mu}_{MA}(t) = \frac{\sum_s \pi_k^{-1} Y_k(t)}{\sum_s \pi_k^{-1}}, \quad t \in [0, T].$$

Lorsque la taille de l'échantillon est variable ou lorsque les Y_k sont très peu corrélés aux π_k, il peut être préférable d'utiliser l'estimateur de Hájek plutôt que l'estimateur de Horvitz-Thompson $\hat{\mu}$ (pour plus de détails se référer à Särndal et al. *(1992), Chapitre 5.7).*

Remarque 2.2. *L'estimateur $\widehat{\mu}_{MA}$ peut également être obtenu à l'aide d'une approche de type calage (Deville et Särndal (1992)). Par exemple, si nous considérons la distance de type khi-2, nous obtenons les nouveaux poids suivants :*

$$w_{ks} = \frac{1}{\pi_k} - \left(\sum_s \frac{\mathbf{x}_l}{\pi_l} - \sum_U \mathbf{x}_l \right)' \left(\sum_s \frac{\mathbf{x}_l \mathbf{x}_l'}{\pi_l} \right)^{-1} \frac{\mathbf{x}_k}{\pi_k}.$$

Dans ce cas, l'estimateur par calage $\sum_s w_{ks} Y_k(t)/N$ de la moyenne $\mu(t)$ est égal à $\widehat{\mu}_{MA}(t)$ défini dans l'équation (2.5).

D'autres distances ont été proposées dans Deville et Särndal (1992). On peut montrer que, pour certaines distances, l'estimateur par calage de la moyenne $\mu(t)$ est ponctuellement asymptotiquement équivalent à l'estimateur $\widehat{\mu}_{MA}(t)$.

2.1.1 Un estimateur régularisé pour le cadre asymptotique

La construction de l'estimateur $\hat{\mu}_{MA}(t)$ repose sur l'hypothèse que la matrice $\widehat{\mathbf{G}}$ est inversible. Pour montrer la convergence uniforme, nous devons introduire une modification de $\widehat{\mathbf{G}}$ qui permettra de contrôler la norme de son inverse. Cette astuce a déjà été utilisée dans Bosq (2000) et Guillas (2001) pour l'estimation des processus autorégressifs fonctionnels. La matrice $\widehat{\mathbf{G}}$ étant une matrice non négative et symétrique de taille $p \times p$, elle peut s'écrire de la manière suivante

$$\widehat{\mathbf{G}} = \sum_{j=1}^{p} \lambda_{j,n} \mathbf{v}_{jn} \mathbf{v}_{jn}',$$

où $\mathbf{v}_{jn}$ est le vecteur propre orthonormé associé à la j-ème valeur propre $\lambda_{j,n}$, avec $\lambda_{1,n} \geq .. \geq \lambda_{p,n} \geq 0$. Soit un réel $a > 0$. L'estimateur régularisé de $\mathbf{G}$ est défini par

$$\widehat{\mathbf{G}}_a = \sum_{j=1}^{p} \max(\lambda_{j,n}, a)\ \mathbf{v}_{jn}\mathbf{v}'_{jn}.$$

Remarque 2.3.

i. L'estimateur $\widehat{\mathbf{G}}_a$ est toujours inversible.

ii. Si $\lambda_{p,n} \geq a$ alors $\widehat{\mathbf{G}} = \widehat{\mathbf{G}}_a$.

iii.

$$\|\widehat{\mathbf{G}}_a^{-1}\| \leq a^{-1}, \tag{2.7}$$

où $\|.\|$ désigne la norme matricielle classique.

Si $a > 0$ est suffisamment petit, nous pouvons montrer que sous certaines conditions sur les moments des variables $X_1, \ldots, X_p$ et sur les probabilités d'inclusion d'ordre un et deux que $\mathbb{P}(\lambda_{p,n} \leq a) = O(n^{-1})$ (cf. Lemme 2.1 dans la section 2.4.1).

Il est alors possible d'estimer la courbe $\mu(t)$ en remplaçant $\widehat{\mathbf{G}}$ par $\widehat{\mathbf{G}}_a$ dans la formule (2.5). La courbe moyenne est alors estimée par

$$\widehat{\mu}_{MA,a}(t) = \frac{1}{N}\sum_{k\in U}\widehat{Y}_{k,a}(t) - \frac{1}{N}\sum_{k\in s}\frac{\widehat{Y}_{k,a}(t) - Y_k(t)}{\pi_k}, \quad t \in [0,T], \tag{2.8}$$

où $\widehat{Y}_{k,a}(t) = \mathbf{x}'_k\widehat{\boldsymbol{\beta}}_a(t)$ et $\widehat{\boldsymbol{\beta}}_a(t) = \widehat{\mathbf{G}}_a^{-1}\frac{1}{N}\sum_s \frac{\mathbf{x}_k Y_k(t)}{\pi_k}$.

2.1.2 Estimation à l'aide de trajectoires discrétisées

Dans la pratique, les trajectoires $Y_k(t)$ ne sont pas observées continument pour $t \in [0,T]$, mais uniquement sur un ensemble de D instants de mesure $0 = t_1 \leq ... \leq t_D = T$. Si on suppose que les trajectoires sont suffisamment régulières et qu'il n'y a pas d'erreur de mesure, nous pouvons utiliser une interpolation linéaire pour approximer les courbes à chaque instant t (cf. section 1.7.1, Cardot et Josserand (2011) et Cardot *et al.* (2012)). Pour chaque unité k de l'échantillon s, la courbe interpolée est définie par

$$Y_{k,d}(t) = Y_k(t_i) + \frac{Y_k(t_{i+1}) - Y_k(t_i)}{t_{i+1} - t_i}(t - t_i), \quad t \in [t_i, t_{i+1}], \tag{2.9}$$

et l'estimateur $\widehat{\boldsymbol{\beta}}_{a,d}(t)$ de $\boldsymbol{\beta}(t)$ basé sur les courbes interpolées est défini comme suit

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{a,d}(t) &= \widehat{\mathbf{G}}_a^{-1}\frac{1}{N}\sum_s \mathbf{x}_k Y_{k,d}(t) \\ &= \widehat{\boldsymbol{\beta}}_a(t_i) + \frac{\widehat{\boldsymbol{\beta}}_a(t_{i+1}) - \widehat{\boldsymbol{\beta}}_a(t_i)}{t_{i+1} - t_i}(t - t_i), \quad t \in [t_i, t_{i+1}].\end{aligned} \tag{2.10}$$

Dans ce cas, l'estimateur de la courbe moyenne basé sur les observations discrétisées est obtenu en faisant une interpolation linéaire entre les estimateurs $\widehat{\mu}_{MA,a}(t_i)$ et

$\widehat{\mu}_{MA,a}(t_{i+1})$,

$$\begin{aligned}\widehat{\mu}_{MA,d}(t) &= \widehat{\mu}_{MA,a}(t_i) + \frac{\widehat{\mu}_{MA,a}(t_{i+1}) - \widehat{\mu}_{MA,a}(t_i)}{t_{i+1} - t_i}(t - t_i) \\ &= \frac{1}{N}\sum_{k\in U}\widehat{Y}_{k,d}(t) - \frac{1}{N}\sum_{k\in s}\frac{(\widehat{Y}_{k,d}(t) - Y_{k,d}(t))}{\pi_k}, \end{aligned} \tag{2.11}$$

où $\widehat{Y}_{k,d}(t) = \mathbf{x}_k'\widehat{\boldsymbol{\beta}}_{a,d}(t)$ et $t \in [t_i, t_{i+1}]$.

2.2 Propriétés asymptotiques sous le plan de sondage

2.2.1 Hypothèses

Pour montrer les propriétés asymptotiques, sous le plan de sondage $p(.)$, de l'estimateur $\widehat{\mu}_{MA,d}$, nous devons supposer que la taille de l'échantilllon et celle de la population deviennent très larges. Nous nous plaçons alors dans le cadre asymptotique introduit par Isaki et Fuller (1982) et décrit dans la section 1.5. Nous avons également besoin des hypothèses suivantes

B1. Nous supposons que $\lim_{N\to\infty}\frac{n}{N} = \pi \in]0,1[$.

B2. Nous supposons que $\min_{k\in U}\pi_k \geq \lambda > 0$, $\min_{k\neq l}\pi_{kl} \geq \lambda^* > 0$ et qu'il existe une constante C_1 telle que

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

B3. Il existe deux constantes positives C_2 et C_3 et $\beta > 1/2$ telles que, pour tout N et pour tout $(r,t) \in [0,T]\times[0,T]$,

$$\frac{1}{N}\sum_{k\in U} Y_k(0)^2 < C_2 \quad \text{et} \quad \frac{1}{N}\sum_{k\in U}\{Y_k(t) - Y_k(r)\}^2 < C_3|t-r|^{2\beta}.$$

B4. Il existe une constante positive C_4 telle que pour tout $k \in U$, $\|\mathbf{x}_k\|^2 < C_4$.

B5. Nous supposons que, pour $N > N_0$, la matrice $\mathbf{G}$ est inversible et que le nombre a choisi précédemment satisfait $\|\mathbf{G}^{-1}\| < a^{-1}$.

Les hypothèses **B1** et **B2** ont déjà été introduites par Breidt et Opsomer (2000). La première suppose que la taille de l'échantillon et celle de la population augmentent à la même vitesse. La seconde hypothèse quantifie l'écart à l'indépendance des probabilités d'inclusion d'ordre deux et permet de donner la vitesse de convergence. Ces deux hypothèses sont satisfaites pour quelques plans de sondage à taille fixe (ex : sondage aléatoire simple sans remise, le tirage réjectif). Pour de plus amples détails se référer à Hájek (1981), Robinson et Särndal (1983) et Breidt et Opsomer (2000).

L'hypothèse **B3** est une condition minimale de régularité déjà introduite par Cardot et Josserand (2011). Même si la convergence ponctuelle, pour chaque valeur de t fixée, peut être prouvée sans condition sur le coefficient de Hölder β, cette condition est nécessaire pour obtenir la convergence uniforme de l'estimateur $\widehat{\mu}_{MA,d}$ vers μ. Cette

hypothèse implique également que les trajectoires des processus résiduels ϵ_{kt}, définis dans l'équation (2.2), sont assez régulières mais pas nécessairement dérivables.

L'hypothèse **B4** pourrait certainement être affaiblie mais cela serait au détriment de la longueur des démonstrations.

L'hypothèse **B5** signifie que pour tout $\mathbf{u} \in \mathbb{R}^p$, avec $\mathbf{u} \neq 0$, nous avons $\mathbf{u}'\mathbf{G}\mathbf{u} \geq a\mathbf{u}'\mathbf{u}$. Isaki et Fuller (1982) ont utilisé le même type d'hypothèse pour montrer la convergence en probabilité ponctuelle. Robinson et Särndal (1983) ont introduit une condition plus forte (condition A7 dans leur article) liée à la convergence quadratique de l'estimateur du vecteur des coefficients de régression $\boldsymbol{\beta}$.

2.2.2 Convergence uniforme

Sous des conditions supplémentaires sur le schéma de discrétisation, nous pouvons montrer la convergence uniforme de l'estimateur $\widehat{\boldsymbol{\beta}}_{a,d}(t)$ vers $\tilde{\boldsymbol{\beta}}(t)$.

Proposition 2.1. *Supposons que les hypothèses* **B1-B5** *sont vérifiées. Si le schéma de discrétisation satisfait* $\max_{i=\{1,..,D_N-1\}} |t_{i+1} - t_i|^{2\beta} = o(n^{-1})$ *alors il existe une constante* $C > 0$ *telle que, pour tout n,*

$$\sqrt{n}\ \mathbb{E}_p \left\{ \sup_{t\in[0,T]} \left\| \widehat{\boldsymbol{\beta}}_{a,d}(t) - \tilde{\boldsymbol{\beta}}(t) \right\| \right\} \leq C.$$

Nous pouvons également donner une propriété similaire pour l'estimateur $\widehat{\mu}_{MA,d}$.

Proposition 2.2. *Supposons que les hypothèses* **B1-B5** *sont vérifiées. Si le schéma de discrétisation satisfait* $\max_{i=\{1,..,D_N-1\}} |t_{i+1} - t_i|^{2\beta} = o(n^{-1})$ *alors il existe une constante* $C > 0$ *telle que, pour tout n,*

$$\sqrt{n}\ \mathbb{E}_p \left\{ \sup_{t\in[0,T]} |\ \widehat{\mu}_{MA,d}(t) - \mu(t)\ | \right\} \leq C.$$

Nous déduisons de la Proposition 2.2 que l'estimateur $\widehat{\mu}_{MA,d}(t)$ et asymptotiquement non-biaisé et convergent par rapport au plan de sondage. Nous pouvons également remarquer que, sous l'hypothèse supplémentaire sur les points de discrétisation, l'erreur d'approximation est négligeable par rapport à celle due au plan de sondage. En effet, pour chaque $t \in [0,T]$,

$$\widehat{\mu}_{MA,a}(t) - \widetilde{\mu}(t) \quad = \quad \frac{1}{N} \sum_{k\in U} \left(1 - \frac{\mathbb{1}_k}{\pi_k}\right) \mathbf{x}_k' \left(\widehat{\boldsymbol{\beta}}_a(t) - \widetilde{\boldsymbol{\beta}}(t)\right), \tag{2.12}$$

ainsi, nous pouvons facilement montrer en utilisant les hypothèses précédentes et le Lemme 2.4 que

$$\sqrt{n}\left(\widehat{\mu}_{MA,d}(t) - \widetilde{\mu}(t)\right) = o_p(1), \quad t \in [0,T]. \tag{2.13}$$

2.2.3 Estimation de la fonction de covariance sous le plan de sondage

Dans cette section, nous allons estimer la fonction de covariance, sous le plan de sondage $p(.)$, de l'estimateur $\widehat{\mu}_{MA,d}$ puis nous donnerons ses propriétés asymptotiques.

L'estimateur basé sur le modèle de régression $\widehat{\mu}_{MA,d}$ peut également s'écrire sous la forme d'une combinaison non linéaire d'estimateur de Horvitz-Thompson :

$$\widehat{\mu}_{MA,d} = \frac{1}{N}\sum_s \frac{Y_{k,d}(t)}{\pi_k} + \left(\frac{1}{N}\sum_U \mathbf{x}_k' - \frac{1}{N}\sum_s \frac{\mathbf{x}_k'}{\pi_k}\right)\widehat{\boldsymbol{\beta}}_{a,d}(t), \quad t \in [0,T]. \tag{2.14}$$

Dans ce cas, nous ne pouvons plus utiliser la formule usuelle de la covariance de Horvitz-Thompson (cf. équation (1.47)). Cependant l'erreur d'approximation étant négligeable (cf. équation (2.13)), nous pouvons approximer la fonction de covariance de l'estimateur $\widehat{\mu}_{MA,d}$ entre deux instants r et t par la covariance de l'estimateur par la différence $\mathrm{Cov}_p(\tilde{\mu}(r), \tilde{\mu}(t))$ (cf Särndal *et al.* (1992)). Ainsi la covariance γ_{MA} de $\widehat{\mu}_{MA,d}$ est approximée par la covariance de Horvitz-Thompson appliquée aux résidus $Y_k - \tilde{Y}_k$

$$\begin{aligned}\gamma_{\mathrm{MA}}(r,t) &\approx \frac{1}{N^2}\mathrm{Cov}_p\left(\sum_{k\in s}\frac{Y_k(r)-\tilde{Y}_k(r)}{\pi_k}, \sum_{k\in s}\frac{Y_k(t)-\tilde{Y}_k(t)}{\pi_k}\right)\\ &= \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}(\pi_{kl}-\pi_k\pi_l)\frac{Y_k(r)-\tilde{Y}_k(r)}{\pi_k}\frac{Y_l(t)-\tilde{Y}_l(t)}{\pi_l}, \quad r,t\in[0,T].\end{aligned} \tag{2.15}$$

Remarque 2.4. *Si les résidus $Y_k(t)-\tilde{Y}_k(t)$ sont petits par rapport à $Y_k(t)$, l'estimateur assisté par le modèle peut être plus performant que l'estimateur de Horvitz-Thompson.*

La fonction de covariance $\gamma_{\mathrm{MA}}(r,t)$ est estimée à l'aide de l'estimateur de la covariance de Horvitz-Thompson appliqué aux résidus estimés $Y_{k,d}(t) - \widehat{Y}_{k,d}(t)$,

$$\widehat{\gamma}_{\mathrm{MA},d}(r,t) = \frac{1}{N^2}\sum_{k,l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\cdot\frac{Y_{k,d}(r)-\widehat{Y}_{k,d}(r)}{\pi_k}\cdot\frac{Y_{l,d}(t)-\widehat{Y}_{l,d}(t)}{\pi_l}, \quad r,t\in[0,T], \tag{2.16}$$

où $\widehat{Y}_{k,d}(t) = \mathbf{x}_k'\widehat{\boldsymbol{\beta}}_{a,d}(t)$.

Pour montrer la convergence de l'estimateur de la covariance $\widehat{\gamma}_{\mathrm{MA},d}(r,t)$, nous devons introduire des hypothèses supplémentaires sur les probabilités d'inclusion d'ordre supérieur et sur les moments d'ordre 4 des trajectoires.

B6. Nous supposons que $\lim_{N\to\infty}\max_{(k,l,k',l')\in D_{4,n}}|E\{(\mathbb{1}_{kl}-\pi_{kl})(\mathbb{1}_{k'l'}-\pi_{k'l'})\}| = 0$ où $D_{t,N}$ est l'ensemble de tous les t-tuples distincts $(i_1,...,i_t)$ dans U.

B7. Il existe deux constantes positives C_5 et C_6 tel que

$$N^{-1}\sum_U Y_k(0)^4 < C_5 \quad \text{et} \quad N^{-1}\sum\{Y_k(t)-Y_k(r)\}^4 < C_6|t-r|^{4\beta},$$

pour tout $(r,t)\in[0,T]^2$.

L'hypothèse **B6** a déjà été posée par Breidt et Opsomer (2000) pour un estimateur assisté par un modèle non-paramétrique et par Cardot et Josserand (2011) pour montrer la convergence de l'estimateur de la covariance de l'estimateur de Horvitz-Thompson. Elle est vérifiée pour des plans simples tels que le sondage aléatoire simple sans remise (SRSWOR), le sondage stratifié avec un SRSWOR à l'intérieur de chaque strate et le sondage réjectif (Breidt et Opsomer (2000), Boistard *et al.* (2012)).

Proposition 2.3. *Supposons que les hypothèses* **B1**-**B7** *sont vérifiées et que le schéma de discrétisation satisfait* $\max_{i=\{1,..,D_N-1\}} |t_{i+1} - t_i| = o(1)$. *Alors, quand* N *tend à l'infini, nous avons pour tout* $(r,t) \in [0,T]^2$,

$$n\ \mathbb{E}_p\left\{\left|\widehat{\gamma}_{MA,d}(r,t) - \gamma_{MA}(r,t)\right|\right\} \quad \to \quad 0$$

et

$$n\ \mathbb{E}_p\left\{\sup_{t\in[0,T]}\left|\widehat{\gamma}_{MA,d}(t,t) - \gamma_{MA}(t,t)\right|\right\} \quad \to \quad 0.$$

Puisque $n\gamma_{\mathrm{MA}}(r,t)$ est borné, nous déduisons de la proposition précédente que l'estimateur de la covariance $\widehat{\gamma}_{\mathrm{MA},d}$ est ponctuellement convergent et la fonction de variance estimée est uniformément convergente.

Remarque 2.5. *Dans la Proposition 2.3, nous ne donnons pas les taux de convergence, la condition sur le nombre de points de discrétisation est donc plus faible que dans la Proposition 2.2. Pour obtenir de tels taux, il faudrait ajouter des hypothèses supplémentaires sur le plan de sondage.*

2.3 Normalité asymptotique

Pour montrer la normalité asymptotique de l'estimateur $\hat{\mu}_{\mathrm{MA},d}$ dans l'espace des fonctions continues, nous devons introduire une nouvelle hypothèse.

B8. Supposons que pour chaque valeur de t fixée, $t \in [0,1]$,

$$\{\gamma_{\mathrm{MA}}(t,t)\}^{-1/2}\left(\tilde{\mu}(t) - \mu(t)\right) \quad \to \quad \mathcal{N}(0,1)$$

en distribution quand N tend à l'infini.

Cette hypothèse est satisfaite pour certains plans de sondage classiques (cf. *e.g.* Fuller (2009a), Chapitre 2.2).

De la relation (2.13), nous déduisons que pour chaque instant t fixé, $t \in [0,T]$,

$$\widehat{\mu}_{\mathrm{MA},d}(t) - \mu(t) = \tilde{\mu}(t) - \mu(t) + o_p(n^{-1/2}),$$

Ainsi, si les conditions de la Proposition 2.2 sont vérifiées, $\sqrt{n}\left(\widehat{\mu}_{\mathrm{MA},d}(t) - \mu(t)\right)$ est également ponctuellement asymptotiquement gaussien.

Proposition 2.4. *Supposons que les hypothèses* **B1-B5** *et* **B8** *sont vérifiées. Si le schéma de discrétisation satisfait* $\max_{i=\{1,..,D_N-1\}} |t_{i+1}-t_i|^{2\beta} = o(n^{-1})$, *nous avons quand n tend à l'infini*

$$\sqrt{n}\left\{\widehat{\mu}_{MA,d}-\mu\right\} \to Z \text{ en distribution dans } C[0,T] \tag{2.17}$$

où Z est un processus gaussien à valeur dans $C[0,T]$ *de moyenne 0 et de fonction de covariance* $\gamma_Z(r,t) = \lim_{n\to+\infty} n\gamma_{MA}(r,t)$.

2.4 Preuves

Tout au long des preuves, nous utiliserons la lettre C pour désigner une constante générique dont la valeur peut varier d'un endroit à l'autre. Définissons $\alpha_k = \frac{\mathbb{1}_k}{\pi_k} - 1$, $\Delta_{kk} = \pi_k(1-\pi_k)$ et $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ pour $k,l \in U$. La norme euclidienne d'un vecteur $\mathbf{v}$ sera notée $\|\mathbf{v}\|$ et la norme d'une matrice $\mathbf{A}$ sera notée $\|\mathbf{A}\| = \sqrt{tr(\mathbf{A}'\mathbf{A})}$.

2.4.1 Quelques lemmes utiles

Lemme 2.1. *Supposons les hypothèses* **B1**, **B2**, **B4** *et* **B5** *vérifiées. Alors, il existe une constante C telle que*

$$n\ \mathbb{E}_p\left(\|\widehat{\mathbf{G}}_a^{-1} - \mathbf{G}^{-1}\|^2\right) \le C.$$

Ce résultat, parfois fixé comme hypothèse (cf. Robinson et Särndal (1983)), nous servira à démontrer la convergence de l'estimateur $\hat{\mu}_{\mathrm{MA},d}$.

Démonstration. La preuve suit les étapes de (Bosq (2000), Théorème 8.4) et (Cardot *et al.* (2010a), Proposition 3.1). Selon l'hypothèse **A5** et l'inégalité (2.7), nous avons

$$\begin{aligned}\|\widehat{\mathbf{G}}_a^{-1} - \mathbf{G}^{-1}\| &\le \|\widehat{\mathbf{G}}_a^{-1}\|.\|\widehat{\mathbf{G}}_a - \mathbf{G}\|.\|\mathbf{G}^{-1}\| \\ &\le a^{-2}\|\widehat{\mathbf{G}}_a - \mathbf{G}\|,\end{aligned}$$

ce qui implique

$$\mathbb{E}_p\left(\|\widehat{\mathbf{G}}_a^{-1} - \mathbf{G}^{-1}\|^2\right) \le a^{-4}\mathbb{E}_p\left(\|\widehat{\mathbf{G}}_a - \mathbf{G}\|^2\right). \tag{2.18}$$

Pour borner $\mathbb{E}_p\left(\|\widehat{\mathbf{G}}_a - \mathbf{G}\|^2\right)$, nous utilisons la décomposition suivante.

$$\begin{aligned}\mathbb{E}_p\left(\|\widehat{\mathbf{G}}_a - \mathbf{G}\|^2\right) &= \mathbb{E}_p\left(\|\widehat{\mathbf{G}}_a - \mathbf{G}\|^2\mathbb{1}_{\{\widehat{\mathbf{G}}_a=\widehat{\mathbf{G}}\}}\right) + \mathbb{E}_p\left(\|\widehat{\mathbf{G}}_a - \mathbf{G}\|^2\mathbb{1}_{\{\widehat{\mathbf{G}}_a\neq\widehat{\mathbf{G}}\}}\right) \\ &\le \mathbb{E}_p\left(\|\widehat{\mathbf{G}} - \mathbf{G}\|^2\right) + 2\mathbb{E}_p\left(\|\widehat{\mathbf{G}}_a - \widehat{\mathbf{G}}\|^2\mathbb{1}_{\{\widehat{\mathbf{G}}_a\neq\widehat{\mathbf{G}}\}}\right) + 2\mathbb{E}_p\left(\|\widehat{\mathbf{G}} - \mathbf{G}\|^2\mathbb{1}_{\{\widehat{\mathbf{G}}_a\neq\widehat{\mathbf{G}}\}}\right) \\ &\le 3\mathbb{E}_p(\|\widehat{\mathbf{G}} - \mathbf{G}\|^2) + 2\mathbb{E}_p\left(\|\widehat{\mathbf{G}}_a - \widehat{\mathbf{G}}\|^2\mathbb{1}_{\{\widehat{\mathbf{G}}_a\neq\widehat{\mathbf{G}}\}}\right).\end{aligned} \tag{2.19}$$

Ensuite nous montrons que (voir aussi la démonstration de la Proposition 3.1 dans Cardot *et al.* (2010a)),

$$\mathbb{E}_p\|\widehat{\mathbf{G}} - \mathbf{G}\|^2 = O(n^{-1}). \tag{2.20}$$

Sous les hypothèses **B1**, **B2** et **B4**,

$$\begin{aligned}
\mathbb{E}_p\|\widehat{\mathbf{G}}-\mathbf{G}\|^2 &= \frac{1}{N^2}\mathbb{E}_p\left(\sum_U\sum_U \alpha_k\alpha_l \mathrm{tr}[\mathbf{x}_k\mathbf{x}_k'\mathbf{x}_l\mathbf{x}_l']\right)\\
&\leq \frac{1}{N^2}\frac{1}{\lambda}\sum_U \mathrm{tr}[\mathbf{x}_k\mathbf{x}_k'\mathbf{x}_k\mathbf{x}_k'] + \max_{k\neq l}|\Delta_{kl}|\frac{1}{N^2\lambda^2}\sum_{k\in U}\sum_{l\in U}\mathrm{tr}[\mathbf{x}_k\mathbf{x}_k'\mathbf{x}_l\mathbf{x}_l']\\
&\leq \frac{1}{N^2}\frac{1}{\lambda}\sum_U \|\mathbf{x}_k\mathbf{x}_k'\|^2 + \max_{k\neq l}|\Delta_{kl}|\frac{1}{N^2\lambda^2}\sum_{k\in U}\sum_{l\in U}\|\mathbf{x}_k\|^2\|\mathbf{x}_l\|^2\\
&\leq \frac{1}{N^2}\frac{1}{\lambda}\sum_U \|\mathbf{x}_k\|^4 + \max_{k\neq l}|\Delta_{kl}|\frac{1}{N\lambda^2}\sum_U\|\mathbf{x}_k\|^4\\
&\leq \frac{1}{n}\left(\frac{n}{N}\frac{1}{\lambda} + n\max_{k\neq l}|\Delta_{kl}|\frac{1}{\lambda^2}\right)C_4^2\\
&\leq \frac{C}{n}.
\end{aligned}$$

De plus

$$\mathbb{E}_p\left(\|\widehat{\mathbf{G}}_a-\widehat{\mathbf{G}}\|^2\mathbb{1}_{\{\widehat{\mathbf{G}}_a\neq\widehat{\mathbf{G}}\}}\right)\leq a^2\mathbb{P}(\widehat{\mathbf{G}}_a\neq\widehat{\mathbf{G}})$$

du fait que

$$\begin{aligned}
\|\widehat{\mathbf{G}}_a-\widehat{\mathbf{G}}\|^2 &= \left\|\sum_{j=1}^p[\max(\lambda_{j,n},a)-\lambda_{j,n}]\mathbf{v}_{jn}\mathbf{v}_{jn}'\right\|^2\\
&\leq \sup_{j=1,\dots,p}|\max(\lambda_{j,n},a)-\lambda_{j,n}|^2\\
&\leq a^2.
\end{aligned}$$

Etant donné que $a<\lambda_p=\left\|\mathbf{G}^{-1}\right\|^{-1}$, nous pouvons utiliser l'inégalité de Chebychev pour borner

$$\begin{aligned}
\mathbb{P}(\widehat{\mathbf{G}}_a\neq\widehat{\mathbf{G}}) &= \mathbb{P}(\lambda_{pn}<a)\\
&\leq \mathbb{P}\left(|\lambda_{pn}-\lambda_p|\geq\frac{|\lambda_p-a|}{2}\right)\\
&\leq \frac{4}{(\lambda_p-a)^2}\mathbb{E}_p\left(|\lambda_{pn}-\lambda_p|^2\right).
\end{aligned}$$

Or la fonction valeur propre est lipschitzienne pour les matrices symétriques. Cela signifie que pour deux matrices $\mathbf{A}$ et $\mathbf{B}$, de valeurs propres respectives $\lambda_1(\mathbf{A})\geq\lambda_2(\mathbf{A})\geq\cdots\geq\lambda_p(\mathbf{A})$ et $\lambda_1(\mathbf{B})\geq\cdots\geq\lambda_p(\mathbf{B})$, nous avons

$$\max_{j\in\{1,\dots,p\}}|\lambda_j(\mathbf{A})-\lambda_j(\mathbf{B})|\leq\|\mathbf{A}-\mathbf{B}\|.$$

Ainsi

$$\mathbb{P}(\widehat{\mathbf{G}}_a\neq\widehat{\mathbf{G}}) \leq \frac{4}{(\lambda_p-a)^2}\mathbb{E}_p\left(\|\widehat{\mathbf{G}}-\mathbf{G}\|^2\right).$$

Par conséquent, pour une certaine constante C,

$$\begin{aligned}
\mathbb{E}_p\left(\|\widehat{\mathbf{G}}_a-\mathbf{G}\|^2\right) &\leq 3\mathbb{E}_p\left(\|\widehat{\mathbf{G}}-\mathbf{G}\|^2\right)+2a^2\mathbb{P}(\widehat{\mathbf{G}}_a\neq\widehat{\mathbf{G}})\\
&\leq \frac{C}{n}. \qquad (2.21)
\end{aligned}$$

Des équations (2.18) et (2.21), nous déduisons le résultat annoncé. □

Lemme 2.2. *Sous les hypothèses* **B1**, **B2** *et* **B4**, *il existe une constante C telle que, pour tout n,*

$$n\ \mathbb{E}_p \left\| \frac{1}{N} \sum_U \left(\frac{\mathbb{1}_k}{\pi_k} - 1 \right) \mathbf{x}_k \right\|^2 \leq C.$$

Démonstration. Sous les hypothèses **B1**, **B2** et **B4**

$$\begin{aligned}
n\mathbb{E}_p \left\| \frac{1}{N} \sum_U \alpha_k \mathbf{x}_k \right\|^2 &= n\mathbb{E}_p \left(\frac{1}{N^2} \sum_{k\in U} \sum_{l\in U} \alpha_k \alpha_l \mathbf{x}_k' \mathbf{x}_l \right) \\
&\leq \frac{n}{N^2} \sum_{k\in U} \sum_{l\in U} \left| \frac{\Delta_{kl}}{\pi_k \pi_l} \right| \mathbf{x}_k' \mathbf{x}_l \\
&\leq \frac{n}{N^2} \left(\frac{1}{\lambda} \sum_{k\in U} \|\mathbf{x}_k\|^2 + \frac{\max_{k\neq l} |\Delta_{kl}|}{\lambda^2} \sum_{\substack{k,l\in U \\ k\neq l}} \|\mathbf{x}_k\| \|\mathbf{x}_l\| \right) \\
&\leq \left[\frac{n}{N} \frac{1}{\lambda} + \frac{1}{\lambda^2} n \max_{k\neq l} |\Delta_{kl}| \right] \frac{1}{N} \sum_{k\in U} \|\mathbf{x}_k\|^2 \\
&\leq \left[\frac{n}{N} \frac{1}{\lambda} + \frac{1}{\lambda^2} n \max_{k\neq l} |\Delta_{kl}| \right] C_4 \\
&\leq C
\end{aligned}$$

pour une certaine constante C. □

Lemme 2.3. *Sous les hypothèses* **B2**-**B5**, *nous avons*

i)

$$\|\tilde{\boldsymbol{\beta}}(t) - \tilde{\boldsymbol{\beta}}(r)\|^2 \leq a^{-2} C_3 C_4 |t - r|^{2\beta}.$$

ii)

$$\|\widehat{\boldsymbol{\beta}}_a(t) - \widehat{\boldsymbol{\beta}}_a(r)\|^2 \leq \frac{a^{-2}}{\lambda^2} C_3 C_4 |t - r|^{2\beta}.$$

Démonstration.

i) Sous les hypothèses **B3**, **B4** et **B5**,

$$\begin{aligned}
\|\tilde{\boldsymbol{\beta}}(t) - \tilde{\boldsymbol{\beta}}(r)\|^2 &= \left\| \mathbf{G}^{-1} \frac{1}{N} \sum_U \mathbf{x}_k (Y_k(t) - Y_k(r) \right\|^2 \\
&\leq \|\mathbf{G}^{-1}\|^2 \left(\frac{1}{N} \sum_U \|\mathbf{x}_k\|^2 \right) \left(\frac{1}{N} \sum_U (Y_k(t) - Y_k(r))^2 \right) \\
&\leq a^{-2} C_4 C_3 |t - r|^{2\beta}.
\end{aligned}$$

ii) La preuve est similaire au point i) mais nécessite de faire une hypothèse supplémentaire pour minorer les probabilités d'inclusion d'ordre un (hypothèse **B2**),

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}_a(t)-\widehat{\boldsymbol{\beta}}_a(r)\|^2 &= \left\|\widehat{\mathbf{G}}_a^{-1}\frac{1}{N}\sum_U \frac{\mathbb{1}_k}{\pi_k}\mathbf{x}_k(Y_k(t)-Y_k(r)\right\|^2 \\
&\leq \frac{1}{\lambda^2}\|\widehat{\mathbf{G}}_a^{-1}\|^2\left(\frac{1}{N}\sum_U\|\mathbf{x}_k\|^2\right)\left(\frac{1}{N}\sum_U(Y_k(t)-Y_k(r))^2\right) \\
&\leq a^{-2}\frac{1}{\lambda^2}C_4C_3|t-r|^{2\beta}.
\end{aligned}$$

□

Le lemme suivant démontre la convergence ponctuelle quadratique moyenne pour toute valeur fixe de $t \in [0,T]$.

Lemme 2.4. *Supposons que les hypothèses* **B1-B5** *sont vérifiées. Alors, il existe une constante positive* ζ_1 *telle que, pour tout* $t \in [0,T]$,

$$n\mathbb{E}_p\left(\|\widehat{\boldsymbol{\beta}}_a(t)-\tilde{\boldsymbol{\beta}}(t)\|^2\right) \leq \zeta_1.$$

Démonstration. La démonstration est similaire à celle du Lemme 2.5 et sera donc omise. □

Lemme 2.5. *Supposons que les hypothèses* **B1-B5** *sont vérifiées. Alors, il existe une constante positive* ζ_2 *telle que,*

$$n\mathbb{E}_p\left(\|\widehat{\boldsymbol{\beta}}_a(t)-\tilde{\boldsymbol{\beta}}(t)-\widehat{\boldsymbol{\beta}}_a(r)+\tilde{\boldsymbol{\beta}}(r)\|^2\right) \leq \zeta_2|t-r|^{2\beta}.$$

Démonstration. Une décomposition directe conduit à

$$\begin{aligned}
& n\|\widehat{\boldsymbol{\beta}}_a(t)-\tilde{\boldsymbol{\beta}}(t)-\widehat{\boldsymbol{\beta}}_a(r)+\tilde{\boldsymbol{\beta}}(r)\|^2 \\
= \; & n\left\|(\widehat{\mathbf{G}}_a^{-1}-\mathbf{G}^{-1})\frac{1}{N}\sum_U\frac{\mathbb{1}_k}{\pi_k}\mathbf{x}_k(Y_k(t)-Y_k(r))+\mathbf{G}^{-1}\frac{1}{N}\sum_U\left(\frac{\mathbb{1}_k}{\pi_k}-1\right)\mathbf{x}_k(Y_k(t)-Y_k(r))\right\|^2 \\
\leq \; & 2n\|\widehat{\mathbf{G}}_a^{-1}-\mathbf{G}^{-1}\|^2\left\|\frac{1}{N}\sum_U\frac{\mathbb{1}_k}{\pi_k}\mathbf{x}_k(Y_k(t)-Y_k(r))\right\|^2 \\
& +2n\|\mathbf{G}^{-1}\|^2\left\|\frac{1}{N}\sum_U\alpha_k\mathbf{x}_k(Y_k(t)-Y_k(r))\right\|^2 \\
:= \; & 2A_{1N}^2+2A_{2N}^2, \qquad (2.22)
\end{aligned}$$

A partir des hypothèses **B2-B4** et de l'inégalité de Cauchy-Schwarz, nous obtenons

$$\begin{aligned}
A_{1N}^2 &\leq n\|\widehat{\mathbf{G}}_a^{-1}-\mathbf{G}^{-1}\|^2\left(\frac{1}{\lambda^2}\frac{1}{N}\sum_U\|\mathbf{x}_k\|^2\right)\left(\frac{1}{N}\sum_U(Y_k(t)-Y_k(r))^2\right) \\
&\leq n\|\widehat{\mathbf{G}}_a^{-1}-\mathbf{G}^{-1}\|^2\frac{1}{\lambda^2}C_3C_4|t-r|^{2\beta}.
\end{aligned}$$

Finalement, en utilisant le Lemme 2.1, nous obtenons

$$\mathbb{E}_p(A_{1N}^2) \quad \leq \quad C|t-r|^{2\beta}, \tag{2.23}$$

pour une certaine constante C.

Sous les hypothèses **B1**-**B5** et en utilisant le même de type de développement que le Lemme 2.2, nous obtenons

$$\begin{aligned}
\mathbb{E}_p(A_{2N}^2) &\leq n\|\mathbf{G}^{-1}\|^2\mathbb{E}_p\left(\left\|\frac{1}{N}\sum_U \alpha_k \mathbf{x}_k(Y_k(t)-Y_k(r))\right\|^2\right) \\
&\leq \left(\frac{n}{N}\frac{1}{\lambda}+\frac{n\max_{k\neq l}|\Delta_{kl}|}{\lambda^2}\right)C_3C_4a^{-2}|t-r|^{2\beta} \\
&\leq C|t-r|^{2\beta} \qquad (2.24)
\end{aligned}$$

pour une certaine constante C. Enfin, on combine (2.22), (2.23) et (2.24) pour obtenir le résultat souhaité.

□

Nous allons maintenant donner quelques lemmes utiles pour démontrer la convergence de la variance.

Lemme 2.6. *Supposons les hypothèses* **B2**-**B5** *et* **B7** *vérifiées. Il existe deux constantes* ζ_4 *et* ζ_5 *telles que*

i.

$$\frac{1}{N}\sum_U \tilde{e}_k(t)^2\tilde{e}_k(r)^2 \leq \zeta_4.$$

ii.

$$\frac{1}{N^2}\sum_U\sum_U \tilde{e}_k(t)^2\tilde{e}_l(r)^2 \leq \zeta_5,$$

où $\tilde{e}_k(t) = Y_k(t) - \tilde{Y}_k(t)$.

Démonstration

i. Nous avons

$$\begin{aligned}
\frac{1}{N}\sum_U \tilde{e}_k^2(t)\tilde{e}_k^2(r) &\leq \frac{4}{N}\sum_U\left(Y_k^2(t)Y_k^2(r)+\tilde{Y}_k^2(r)Y_k^2(t)+\tilde{Y}_k^2(t)Y_k^2(r)+\tilde{Y}_k^2(t)\tilde{Y}_k^2(r)\right) \\
&\leq 4\left[\left(\frac{1}{N}\sum_U Y_k^4(t)\right)^{1/2}\left(\frac{1}{N}\sum_U Y_k^4(r)\right)^{1/2}+\left(\frac{1}{N}\sum_U \tilde{Y}_k^4(t)\right)^{1/2}\left(\frac{1}{N}\sum_U Y_k^4(r)\right)^{1/2}\right. \\
&\left.+\left(\frac{1}{N}\sum_U Y_k^4(t)\right)^{1/2}\left(\frac{1}{N}\sum_U \tilde{Y}_k^4(r)\right)^{1/2}+\left(\frac{1}{N}\sum_U \tilde{Y}_k^4(t)\right)^{1/2}\left(\frac{1}{N}\sum_U \tilde{Y}_k^4(r)\right)^{1/2}\right]
\end{aligned}$$

Or, nous pouvons déduire des hypothèses **B4**, **B5** et **B7** qu'il existe une constante C telle que

$$\frac{1}{N}\sum_U Y_k^4(t) < C$$

et

$$\frac{1}{N}\sum_U \tilde{Y}_k^4(r) < C.$$

Ainsi, pour une certaine constante ζ_4,

$$\frac{1}{N}\sum_U \tilde{e}_k(t)^2\tilde{e}_k(r)^2 \leq \zeta_4.$$

ii. La preuve est similaire à celle du point i.

□

Lemme 2.7. *Supposons les hypothèses* **B2-B5** *et* **B7** *vérifiées. Il existe deux constantes* ζ_6 *et* ζ_7 *telles que*

i.

$$\mathbb{E}_p\left(\frac{1}{N}\sum_k \widehat{\tilde{\phi}}_{k,k}(t,r)^2\right) \leq \zeta_6|t-r|^{2\beta}.$$

ii.

$$E_p\left(\frac{1}{N^2}\sum_{k,l} \widehat{\tilde{\phi}}_{k,l}(t,r)\right)^2 \leq \zeta_7[|t-r|^{2\beta},$$

où $\widehat{\tilde{\phi}}_{k,l}(t,r) = \widehat{\tilde{e}}_k(t)\widehat{\tilde{e}}_l(t) - \widehat{\tilde{e}}_k(r)\widehat{\tilde{e}}_l(r)$ *et* $\widehat{\tilde{e}}_k(t) = \tilde{Y}_k(t) - \widehat{Y}_{k,a}(t)$.

Démonstration

i. Nous avons

$$\begin{aligned}\mathbb{E}_p\left(\frac{1}{N}\sum_k \widehat{\tilde{\phi}}_{k,k}^2(t,r)\right) &\leq 2\left[\left(\frac{1}{N}\sum_U |\tilde{Y}_k(t)-\tilde{Y}_k(r)|^4\right)^{1/2} + \mathbb{E}_p\left(\frac{1}{N}\sum_U |\widehat{Y}_{k,a}(t)-\widehat{Y}_{k,a}(r)|^4\right)^{1/2}\right] \\ &\quad .\mathbb{E}_p\left(\frac{1}{N}\sum_U \left[|\tilde{Y}_k(t)|+|\tilde{Y}_k(r)|+|\widehat{Y}_{k,a}(t)|+|\widehat{Y}_{k,a}(r)|\right]^4\right)^{1/2}.\end{aligned} \tag{2.25}$$

En appliquant le Lemme 2.3, nous avons

$$\mathbb{E}_p\left(\frac{1}{N}\sum_U |\widehat{Y}_{k,a}(t)-\widehat{Y}_{k,a}(r)|^4\right) \leq C_4^2\left(\frac{a^{-2}}{\lambda^2}C_3C_4|t-r|^{2\beta}\right)^2 \tag{2.26}$$

et

$$\frac{1}{N}\sum_U |\tilde{Y}_k(t)-\tilde{Y}_k(r)|^4 \leq C_4^2\left(a^{-2}C_3C_4|t-r|^{2\beta}\right)^2. \tag{2.27}$$

Par conséquent, nous obtenons

$$\mathbb{E}_p\left(\frac{1}{N}\sum_k \widehat{\tilde{\phi}}_{k,k}^2(t,r)\right) \leq \zeta_6|t-r|^{2\beta}.$$

ii. La preuve est similaire à celle du point i.

□

Lemme 2.8. *Supposons que les hypothèses* **B2-B5** *et* **B7** *sont vérifiées. Il existe deux constantes* ζ_8 *et* ζ_9 *telles que*

$$\frac{1}{N}\sum_k \phi_{k,k}^2(t,r) \leq \zeta_8|t-r|^{2\beta}$$

et

$$\left(\frac{1}{N^2}\sum_{k,l}\phi_{k,l}(t,r)\right)^2 \leq \zeta_9|t-r|^{2\beta}$$

où $\phi_{k,l}(t,r) = \tilde{e}_k(t)\tilde{e}_l(t) - \tilde{e}_k(r)\tilde{e}_l(r)$ *and* $\tilde{e}_k(t) = Y_k(t) - \tilde{Y}_k(t)$.

Démonstration. La preuve est similaire à celle du Lemme 2.7 et sera donc omise. □

Lemme 2.9. *Supposons que les hypothèses* **B2-B5** *et* **B7** *sont vérifiées. Il existe deux constantes* ζ_{10} *et* ζ_{11} *telles que*

$$\mathbb{E}_p\left(\frac{1}{N}\sum_k \tilde{\phi}_{k,k}(t,r)^2\right) \leq \zeta_{10}|t-r|^{2\beta}$$

et

$$\mathbb{E}_p\left(\frac{1}{N^2}\sum_{k,l}\tilde{\phi}_{k,l}(t,r)\right)^2 \leq \zeta_{11}|t-r|^{2\beta},$$

où $\tilde{\phi}_{k,l}(t,r) = \tilde{e}_k(t)\widehat{\tilde{e}}_l(t) - \tilde{e}_k(r)\widehat{\tilde{e}}_l(r)$, $\tilde{e}_k(t) = Y_k(t) - \tilde{Y}_k(t)$ *et* $\widehat{\tilde{e}}_k(t) = \tilde{Y}_k(t) - \widehat{Y}_{k,a}(t)$.

Démonstration. La preuve est similaire à celle du Lemme 2.7 et sera donc omise. □

2.4.2 Preuve de la Proposition 2.1 et de la Proposition 2.2

La preuve de la Proposition 2.1 est omise. Elle est analogue à la preuve de la Proposition 2.2, qui est donnée ci-dessous. Les différentes étapes de la preuve sont similaires à celles de la Proposition 1 dans Cardot et Josserand (2011).

Considérons la décomposition suivante, pour $t \in [0,T]$,

$$\sup_{t\in[0,T]}|\widehat{\mu}_{\mathrm{MA},d}(t) - \mu(t)| \leq \sup_{t\in[0,T]}|\widehat{\mu}_{\mathrm{MA},d}(t) - \widehat{\mu}_{\mathrm{MA},a}(t)| + \sup_{t\in[0,T]}|\widehat{\mu}_{\mathrm{MA},a}(t) - \mu(t)| \quad (2.28)$$

et étudions séparément chaque terme à droite de l'inégalité.

Etape 1. Erreur d'interpolation $\sup_{t\in[0,T]}|\widehat{\mu}_{\mathrm{MA},d}(t) - \widehat{\mu}_{\mathrm{MA},a}(t)|$.

Supposons que $t \in [t_i, t_{i+1}[$. Dans ce cas, on a

$$|\widehat{\mu}_{\mathrm{MA},d}(t) - \widehat{\mu}_{\mathrm{MA},a}(t)| \leq |\widehat{\mu}_{\mathrm{MA},a}(t_i) - \widehat{\mu}_{\mathrm{MA},a}(t)| + |\widehat{\mu}_{\mathrm{MA},a}(t_{i+1}) - \widehat{\mu}_{\mathrm{MA},a}(t_i)|. \quad (2.29)$$

En utilisant les hypothèses **B2-B5** et le Lemme 2.3, ii), on a

$$\begin{aligned}
|\widehat{\mu}_{\mathrm{MA},a}(t) - \widehat{\mu}_{\mathrm{MA},a}(r)| &\leq \left|\frac{1}{N}\sum_U \alpha_k \mathbf{x}_k'(\widehat{\boldsymbol{\beta}}_a(t) - \widehat{\boldsymbol{\beta}}_a(r))\right| + \frac{1}{N}\sum_s \frac{|Y_k(t) - Y_k(r)|}{\pi_k} \\
&\leq \left(1 + \frac{1}{\lambda}\right)\sqrt{C_4}\|\widehat{\boldsymbol{\beta}}_a(t) - \widehat{\boldsymbol{\beta}}_a(r)\| + \frac{1}{\lambda}\left(\frac{1}{N}\sum_U (Y_k(t) - Y_k(r))^2\right)^{1/2} \\
&\leq \left((1+\lambda^{-1})C_4 a^{-1} + 1\right)\lambda^{-1}\sqrt{C_3}|t-r|^{\beta}.
\end{aligned}$$

Ainsi, il existe une constante positive C telle que

$$|\widehat{\mu}_{\mathrm{MA},a}(t) - \widehat{\mu}_{\mathrm{MA},a}(r)| \leq C|t-r|^{\beta}$$

et par conséquent,

$$\begin{aligned}
|\widehat{\mu}_{\mathrm{MA},d}(t) - \widehat{\mu}_{\mathrm{MA},a}(t)| &\leq C[|t_i - t|^{\beta} + |t_{i+1} - t_i|^{\beta}] \\
&\leq 2C|t_{i+1} - t_i|^{\beta}.
\end{aligned}$$

Par hypothèse, $\max_{i=\{1,\ldots,D_{N-1}\}} |t_{i+1} - t_i|^{\beta} = o(n^{-1/2})$. Nous en déduisons que

$$\sup_{t\in[0,T]} \sqrt{n}|\widehat{\mu}_{\mathrm{MA},d}(t) - \widehat{\mu}_{\mathrm{MA},a}(t)| = o(1). \tag{2.30}$$

Etape 2. Erreur d'estimation $\sup_{t\in[0,T]} |\widehat{\mu}_{\mathrm{MA},a}(t) - \mu(t)|$.

Nous utilisons la décomposition suivante :

$$\sup_{t\in[0,T]} |\widehat{\mu}_{\mathrm{MA},a}(t) - \mu(t)| \leq |\widehat{\mu}_{\mathrm{MA},a}(0) - \mu(0)| + \sup_{r,t\in[0,T]} |\widehat{\mu}_{\mathrm{MA},a}(t) - \mu(t) - \widehat{\mu}_{\mathrm{MA},a}(r) + \mu(r)|. \tag{2.31}$$

Considérons tout d'abord,

$$\begin{aligned}
\widehat{\mu}_{\mathrm{MA},a}(0) - \mu(0) &= \frac{1}{N}\sum_U \alpha_k Y_k(0) - \frac{1}{N}\sum_U \alpha_k \widehat{Y}_k(0) \\
&= \frac{1}{N}\sum_U \alpha_k Y_k(0) - \frac{1}{N^2}\sum_U \alpha_k \mathbf{x}_k' \widehat{\mathbf{G}}_a^{-1} \sum_s \frac{\mathbf{x}_l Y_l(0)}{\pi_l}.
\end{aligned}$$

En utilisant les hypothèses **B1-B3** et le Lemme 2.2, nous obtenons directement que, pour une certaine constante C,

$$\mathbb{E}_p\left(\widehat{\mu}_{\mathrm{MA},a}(0) - \mu(0)\right)^2 \leq \frac{C}{n}. \tag{2.32}$$

En effet,

$$\begin{aligned}
\mathbb{E}_p\left(\widehat{\mu}_{\mathrm{MA},a}(0)-\mu(0)\right)^2 &\leq \frac{2}{n}\left(\frac{n}{N}\frac{1}{\lambda}+\frac{1}{\lambda^2}n\max_{k\neq l}|\Delta_{kl}|\right)\left(\frac{1}{N}\sum_U Y_k(0)^2\right)\\
&\quad +2\frac{1}{\lambda^2N^2}\sum_U\|\mathbf{x}_l\|^2|Y_l(0)|^2\mathbb{E}_p\left(\left\|\frac{1}{N}\sum_U\alpha_k\mathbf{x}_k\right\|^2\|\hat{\mathbf{G}}_a^{-1}\|^2\right)\\
&\leq 2\frac{C_2}{n}\left(\frac{n}{N}\frac{1}{\lambda}+\frac{1}{\lambda^2}n\max_{k\neq l}|\Delta_{kl}|\right)\left(1+\frac{a^{-2}}{\lambda^2}C_4^2\right)\\
&:= \frac{C}{n}. \qquad (2.33)
\end{aligned}$$

Pour majorer le second terme de l'inégalité (2.31), nous utilisons le corollaire 2.2.5 de Van der Vaart et Wellner (2000), basé sur les inégalités maximales. Considérons la norme d'Orlicz définie, pour une variable aléatoire X, par

$$\|X\|_\psi = \sqrt{\mathbb{E}_p(\psi(X))}.$$

Dans le cas particulier où $\psi(u) = u^2$, la norme d'Orlicz est tout simplement égale à la norme L^2, $\|X\|_\psi = \sqrt{\mathbb{E}_p(X^2)}$. Nous introduisons, pour tout $(r,t) \in [0,T]^2$, la semi-métrique $d(r,t)$ définie par

$$\begin{aligned}
d^2(r,t) &= \left\|\sqrt{n}|\widehat{\mu}_{\mathrm{MA},a}(t)-\mu(t)-\widehat{\mu}_{\mathrm{MA},a}(r)-\mu(r)|\right\|_\psi^2\\
&= n\mathbb{E}_p\left(|\widehat{\mu}_{\mathrm{MA},a}(t)-\mu(t)-\widehat{\mu}_{\mathrm{MA},a}(r)+\mu(r)|^2\right)
\end{aligned}$$

Nous considérons le *packing number* $D(\epsilon,d)$ défini comme étant le nombre maximum de points dans $[0,T]$ dont la distance d entre chaque paire est strictement plus grande que ϵ. D'après le Corollaire 2.2.5 dans Van der Vaart et Wellner (2000), il existe une constante $K > 0$ telle que

$$\left\|\sup_{(r,t)\in[0,T]^2}\sqrt{n}|\widehat{\mu}_{\mathrm{MA},a}(t)-\mu(t)-\widehat{\mu}_{\mathrm{MA},a}(r)-\mu(r)|\right\|_\psi \leq K\int_0^T\psi^{-1}(D(\epsilon,d))d\epsilon. \qquad (2.34)$$

Nous montrons ci-dessous qu'il existe une constante C telle que $d^2(r,t) \leq C|t-r|^{2\beta}$. De ce résultat, nous pourrons déduire que l'intégrale à droite de l'inégalité (2.34) est finie lorsque $\beta > 1/2$.

On a

$$d^2(r,t) \leq 2d_1^2(r,t)+2d_2^2(r,t) \qquad (2.35)$$

où

$$d_1^2(r,t) = n\mathbb{E}_p(|\widehat{\mu}_{\mathrm{MA},a}(t)-\tilde{\mu}(t)-\widehat{\mu}_{\mathrm{MA},a}(r)+\tilde{\mu}(r)|^2) \qquad (2.36)$$

et

$$d_2^2(r,t) = n\mathbb{E}_p(|\tilde{\mu}(t)-\mu(t)-\tilde{\mu}(r)+\mu(r)|^2). \qquad (2.37)$$

En utilisant les hypothèses **B2-B4** ainsi que le Lemme 2.5, nous obtenons, pour une certaine constante C,

$$\begin{aligned} d_1^2(r,t) &\leq \mathbb{E}_p\left\{n\left\|\frac{1}{N}\sum_U \alpha_k \mathbf{x}_k\right\|^2 \|\widehat{\boldsymbol{\beta}}_a(t)-\tilde{\boldsymbol{\beta}}(t)-\widehat{\boldsymbol{\beta}}_a(r)+\tilde{\boldsymbol{\beta}}(r)\|^2\right\} \\ &\leq \left(1+\frac{1}{\lambda}\right)^2 \frac{1}{N}\sum_U \|\mathbf{x}_k\|^2 \mathbb{E}_p\left(n\|\widehat{\boldsymbol{\beta}}_a(t)-\tilde{\boldsymbol{\beta}}(t)-\widehat{\boldsymbol{\beta}}_a(r)+\tilde{\boldsymbol{\beta}}(r)\|^2\right) \\ &\leq \left(1+\frac{1}{\lambda}\right)^2 C_4\zeta_2|t-r|^{2\beta} \\ &:= C|t-r|^{2\beta}. \end{aligned} \tag{2.38}$$

Nous avons également

$$\begin{aligned} d_2^2(r,t) &= n\mathbb{E}_p\left[\frac{1}{N}\sum_U \alpha_k\left[Y_k(t)-Y_k(r)-\mathbf{x}_k'(\tilde{\boldsymbol{\beta}}(t)-\tilde{\boldsymbol{\beta}}(r))\right]\right]^2 \\ &\leq 2\mathbb{E}_p(A_N^2)+2\mathbb{E}_p(B_N^2) \end{aligned} \tag{2.39}$$

avec $A_N^2 = n\left(\frac{1}{N}\sum_U \alpha_k\left[Y_k(t)-Y_k(r)\right]\right)^2$ et $B_N^2 = n\left(\frac{1}{N}\sum_U \alpha_k \mathbf{x}_k'(\tilde{\boldsymbol{\beta}}(t)-\tilde{\boldsymbol{\beta}}(r))\right)^2$.

Sous les hypothèses **B1-B3**, nous montrons facilement qu'il existe une constante positive C telle que

$$\mathbb{E}_p(A_N^2) \leq C|t-r|^{2\beta} \tag{2.40}$$

et en appliquant le Lemme 2.2 et le Lemme 2.3, nous pouvons majorer

$$\begin{aligned} \mathbb{E}_p(B_N^2) &\leq \mathbb{E}_p\left[n\left\|\frac{1}{N}\sum_U \alpha_k \mathbf{x}_k\right\|^2\right]\|\tilde{\boldsymbol{\beta}}(t)-\tilde{\boldsymbol{\beta}}(r)\|^2 \\ &\leq C|t-r|^{2\beta}. \end{aligned} \tag{2.41}$$

En combinant les inégalités (2.40) et (2.41) avec (2.35) et (2.38), nous obtenons que

$$d^2(r,t) \leq C|t-r|^{2\beta}, \tag{2.42}$$

pour une certaine constante C.

De l'équation (2.42), nous déduisons que le *Packing Number* $D(\epsilon,d) = O(\epsilon^{-1/\beta})$. Par conséquent,

$$\int_0^T \psi^{-1}(D(\epsilon,d))d\epsilon \leq \int_0^T \epsilon^{\frac{-1}{2\beta}} < \infty \quad \text{quand } \beta > 1/2 \tag{2.43}$$

et ainsi, pour une certaine constante C,

$$\mathbb{E}_p\left(\sqrt{n}\sup_{r,t\epsilon[0,T]}|\widehat{\mu}_{\mathrm{MA},a}(t)-\mu(t)-\widehat{\mu}_{\mathrm{MA},a}(r)+\mu(r)|\right) < C \tag{2.44}$$

En insérant (2.32) et (2.44) dans (2.31), nous obtenons que

$$\sqrt{n}\mathbb{E}_p\left(\sup_{t\epsilon[0,T]}|\widehat{\mu}_{\mathrm{MA},a}(t)-\mu(t)|\right) < C. \tag{2.45}$$

2.4.3 Preuve de la convergence de la fonction de covariance

Tout d'abord nous démontrons que pour tout $(r,t) \in [0,T]^2$, l'estimateur $\widehat{\gamma}_{\mathrm{MA},d}(r,t)$ de la fonction de covariance converge vers $\gamma_{\mathrm{MA}}(r,t)$.

Ensuite, pour prouver la convergence uniforme de l'estimateur de la variance $\widehat{\gamma}_{\mathrm{MA},d}(t,t)$ nous montrons que la variable aléatoire $n(\widehat{\gamma}_{\mathrm{MA},d}(t,t) - \gamma_{\mathrm{MA}}(t,t))$ converge en distribution vers 0 dans l'espace des fonctions continues.

La preuve est décomposée en deux étapes (cf. Billingsley (1968)). Nous montrons d'abord que le vecteur $n\left(\widehat{\gamma}_{\mathrm{MA},a}(t_1,t_1) - \gamma_{\mathrm{MA}}(t_1,t_1), \ldots, \widehat{\gamma}_{\mathrm{MA},a}(t_p,t_p) - \gamma_{\mathrm{MA}}(t_p,t_p)\right)$ converge en distribution vers 0 et ensuite nous montrons que la suite est tendue.

Etape 1. Convergence ponctuelle

Nous voulons prouver que, pour chaque $(t,r) \in [0,T]^2$, nous avons

$$n\mathbb{E}_p\left\{|\,\widehat{\gamma}_{\mathrm{MA},d}(r,t) - \gamma_{\mathrm{MA}}(r,t)\,|\right\} \to 0, \qquad \text{quand } N \to \infty.$$

Considérons la décomposition suivante

$$n(\widehat{\gamma}_{\mathrm{MA},d}(r,t) - \gamma_{\mathrm{MA}}(r,t)) \quad = \quad n(\widehat{\gamma}_{\mathrm{MA},d}(r,t) - \widehat{\gamma}_{\mathrm{MA},a}(r,t)) + n(\widehat{\gamma}_{\mathrm{MA},a}(r,t) - \gamma_{\mathrm{MA}}(r,t))$$

où $\widehat{\gamma}_{\mathrm{MA},a}(r,t)$ est définie par

$$\widehat{\gamma}_{\mathrm{MA},a}(r,t) = \frac{1}{N^2}\sum_{k,l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{Y_k(r) - \widehat{Y}_{k,a}(r)}{\pi_k}\cdot\frac{Y_l(t) - \widehat{Y}_{l,a}(t)}{\pi_l}$$

Nous étudions séparément les erreurs d'interpolation et d'estimation.

Erreur d'interpolation

Supposons que $t \in [t_i, t_{i+1}[$ et $r \in [t_{i'}, t_{i'+1}[$. Nous avons

$$n(\widehat{\gamma}_{\mathrm{MA},d}(r,t) - \widehat{\gamma}_{\mathrm{MA},a}(r,t)) \le A + B,$$

avec

$$\begin{aligned}
A \quad = \quad & \frac{n}{N^2}\sum_{k,l\in U}\frac{|\Delta_{kl}|}{\pi_{kl}\pi_k\pi_l}|(Y_{k,d}(r) - Y_k(r))(Y_{l,d}(t) - Y_l(t)) \\
& +(Y_{k,d}(r) - Y_k(r))(Y_l(t) - \widehat{Y}_{l,d}(t)) + (Y_k(r) - \widehat{Y}_{k,d}(r))(Y_{l,d}(t) - Y_l(t))| \\
\le \quad & \left(\frac{n}{N}\frac{1}{\lambda^2} + n\max_{k\neq l}|\Delta_{kl}|\frac{1}{\lambda^*\lambda^2}\right)\left[\left(\frac{1}{N}\sum_{k\in U}(Y_{k,d}(r) - Y_k(r))^2\frac{1}{N}\sum_{k\in U}(Y_{k,d}(t) - Y_k(t))^2\right)^{1/2}\right. \\
& +\left(\frac{1}{N}\sum_{k\in U}(Y_{k,d}(r) - Y_k(r))^2\frac{1}{N}\sum_{k\in U}(Y_k(t) - \widehat{Y}_{k,d}(t))^2\right)^{1/2} \\
& \left.+\left(\frac{1}{N}\sum_{k\in U}(Y_k(r) - \widehat{Y}_{k,d}(r))^2\frac{1}{N}\sum_{k\in U}(Y_{k,d}(t) - Y_k(t))^2\right)^{1/2}\right]
\end{aligned}$$

et

$$
\begin{aligned}
B &= \frac{n}{N^2}\sum_{k,l\in U}\frac{|\Delta_{kl}|}{\pi_{kl}\pi_k\pi_l}\left|\left(Y_k(r)-\widehat{Y}_{k,d}(r)\right)\left(Y_l(t)-\widehat{Y}_{l,d}(t)\right)-\left(Y_k(r)-\widehat{Y}_{k,a}(r)\right)\left(Y_l(t)-\widehat{Y}_{l,a}(t)\right)\right| \\
&= \frac{n}{N^2}\sum_{k,l\in U}\frac{|\Delta_{kl}|}{\pi_{kl}\pi_k\pi_l}\left|(Y_k(r)-\widehat{Y}_{k,a}(r))(\widehat{Y}_{l,a}(t)-\widehat{Y}_{l,d}(t))+(Y_l(t)-\widehat{Y}_{l,d}(t))(\widehat{Y}_{k,a}(r)-\widehat{Y}_{k,d}(r))\right| \\
&\le \left(\frac{n}{N}\frac{1}{\lambda^2}+n\max_{k\neq l}|\Delta_{kl}|\frac{1}{\lambda^*\lambda^2}\right)\left[\left(\frac{1}{N}\sum_{k\in U}(Y_k(r)-\widehat{Y}_{k,a}(r))^2\frac{1}{N}\sum_{k\in U}(\widehat{Y}_{k,a}(t)-\widehat{Y}_{k,d}(t))^2\right)^{1/2}\right. \\
&\left.+\left(\frac{1}{N}\sum_{k\in U}(Y_k(t)-\widehat{Y}_{k,d}(t))^2\frac{1}{N}\sum_{k\in U}(\widehat{Y}_{k,a}(r)-\widehat{Y}_{k,d}(r))^2\right)^{1/2}\right].
\end{aligned}
$$

Pour $t\in[t_i,t_{i+1}]$, nous avons

$$|Y_{l,d}(t)-Y_l(t)|\le|Y_l(t_i)-Y_l(t)|+|Y_l(t_{i+1})-Y_l(t_i)|$$

et

$$|\widehat{Y}_{l,a}(t)-\widehat{Y}_{l,d}(t)|\le|\widehat{Y}_{l,a}(t)-\widehat{Y}_{l,a}(t_i)|+|\widehat{Y}_{l,a}(t_{i+1})-\widehat{Y}_{l,a}(t_i)|.$$

Or par hypothèses, nous avons $\frac{1}{N}\sum_U(Y_{l,d}(t)-Y_l(t))^2\le C[|t_i-t|^{2\beta}+|t_{i+1}-t_i|^{2\beta}]$, $\frac{1}{N}\sum_U(Y_l(t)-\widehat{Y}_{l,d}(t))^2=O(1)$ et $\frac{1}{N}\sum_U(Y_l(t)-\widehat{Y}_{l,a}(t))^2=O(1)$. Grâce au Lemme 2.3, nous obtenons la majoration suivante

$$|\widehat{Y}_{l,a}(t_i)-\widehat{Y}_{l,a}(t)| \quad\le\quad C_4a^{-1}\frac{1}{\lambda}C_3^{1/2}|t_i-t|^\beta\le C_4a^{-1}\frac{1}{\lambda}C_3^{1/2}|t_{i+1}-t_i|^\beta.$$

Par conséquent, sous l'hypothèse du schéma de discrétisation, nous obtenons que

$$n|\widehat{\gamma}_{\mathrm{MA},d}(r,t)-\widehat{\gamma}_{\mathrm{MA},a}(r,t)|=o(1).$$

Erreur d'estimation

Considérons maintenant,

$$
\begin{aligned}
n(\widehat{\gamma}_{\mathrm{MA},a}(r,t)-\gamma_{\mathrm{MA}}(r,t)) &= \frac{n}{N^2}\sum_k\sum_l\frac{\Delta_{kl}}{\pi_k\pi_l}\left(\frac{\mathbb{1}_{kl}}{\pi_{kl}}-1\right)[Y_k(t)-\tilde{Y}_k(t)][Y_l(r)-\tilde{Y}_l(r)] \\
&\quad+\frac{n}{N^2}\sum_k\sum_l\frac{\Delta_{kl}}{\pi_k\pi_l}\frac{\mathbb{1}_{kl}}{\pi_{kl}}[Y_k(t)-\tilde{Y}_k(t)][\tilde{Y}_l(r)-\widehat{Y}_{l,a}(r)] \\
&\quad+\frac{n}{N^2}\sum_k\sum_l\frac{\Delta_{kl}}{\pi_k\pi_l}\frac{\mathbb{1}_{kl}}{\pi_{kl}}[\tilde{Y}_k(t)-\widehat{Y}_{k,a}(t)][Y_l(r)-\tilde{Y}_l(r)] \\
&\quad+\frac{n}{N^2}\sum_k\sum_l\frac{\Delta_{kl}}{\pi_k\pi_l}\frac{\mathbb{1}_{kl}}{\pi_{kl}}[\tilde{Y}_k(t)-\widehat{Y}_{k,a}(t)][\tilde{Y}_l(r)-\widehat{Y}_{l,a}(r)] \\
&:= A_1(r,t)+A_2(r,t)+A_3(r,t)+A_4(r,t). \qquad (2.46)
\end{aligned}
$$

Définissons $\tilde{e}_k(t)=Y_k(t)-\tilde{Y}_k(t)$ et montrons, dans un premier temps, que

$$\mathbb{E}_p(A_1(r,t)^2)\to 0 \quad \text{quand } N\to\infty.$$

$$
\begin{aligned}
\mathbb{E}_p(A_1(r,t)^2) &= \mathbb{E}_p\left[\frac{n^2}{N^4}\sum_{k,l}\sum_{k',l'}\frac{\Delta_{kl}}{\pi_k\pi_l}\left(\frac{\mathbb{1}_{kl}}{\pi_{kl}}-1\right)\frac{\Delta_{k'l'}}{\pi_{k'}\pi_{l'}}\left(\frac{\mathbb{1}_{k'l'}}{\pi_{k'l'}}-1\right)\tilde{e}_k(t)\tilde{e}_l(r)\tilde{e}_{k'}(t)\tilde{e}_{l'}(r)\right] \\
&= \mathbb{E}_p\left[\frac{n^2}{N^4}\sum_{k}\sum_{k'}\frac{1-\pi_k}{\pi_k}\left(\frac{\mathbb{1}_k}{\pi_k}-1\right)\frac{1-\pi_{k'}}{\pi_{k'}}\left(\frac{\mathbb{1}_{k'}}{\pi_{k'}}-1\right)\tilde{e}_k(t)\tilde{e}_k(r)\tilde{e}_{k'}(t)\tilde{e}_{k'}(r)\right] \\
&+ 2\mathbb{E}_p\left[\frac{n^2}{N^4}\sum_{k}\sum_{\substack{k',l'\\k'\neq l'}}\frac{1-\pi_k}{\pi_k}\left(\frac{\mathbb{1}_k}{\pi_k}-1\right)\frac{\Delta_{k'l'}}{\pi_{k'}\pi_{l'}}\left(\frac{\mathbb{1}_{k'l'}}{\pi_{k'l'}}-1\right)\tilde{e}_k(t)\tilde{e}_k(r)\tilde{e}_{k'}(t)\tilde{e}_{l'}(r)\right] \\
&+ \mathbb{E}_p\left[\frac{n^2}{N^4}\sum_{\substack{k,l\\k\neq l}}\sum_{\substack{k',l'\\k'\neq l'}}\frac{\Delta_{kl}}{\pi_k\pi_l}\left(\frac{\mathbb{1}_{kl}}{\pi_{kl}}-1\right)\frac{\Delta_{k'l'}}{\pi_{k'}\pi_{l'}}\left(\frac{\mathbb{1}_{k'l'}}{\pi_{k'l'}}-1\right)\tilde{e}_k(t)\tilde{e}_l(r)\tilde{e}_{k'}(t)\tilde{e}_{l'}(r)\right] \\
&:= a_1 + a_2 + a_3. \qquad (2.47)
\end{aligned}
$$

Les hypothèses sur les moments des probabilités d'inclusion et le Lemme 2.6 induisent que

$$
a_1 \leq \left(\frac{n^2}{N^3}\frac{1}{\lambda^3}+\frac{n^2}{N^2}\frac{\max_{k\neq k'}|\Delta_{kk'}|}{\lambda^4}\right)\zeta_4
$$

et

$$
a_3 \leq \frac{C}{N}+\frac{(n\max_{k\neq l}|\Delta_{kl}|)^2}{\lambda^4\lambda_*^2}\max_{(k,l,k',l')\in D_{4,n}}|\mathbb{E}_p\{(\mathbb{1}_{kl}-\pi_{kl})(\mathbb{1}_{k'l'}-\pi_{k'l'})\}|\zeta_5
$$

ainsi $a_1 \to 0$ et $a_3 \to 0$ quand $N \to \infty$. En appliquant l'inégalité de Cauchy-Schwarz, nous obtenons également que $a_2 \to 0$ quand $N \to \infty$, et finalement $\mathbb{E}_p(A_1(r,t)^2) \to 0$ quand $N \to \infty$.

Montrons maintenant que $\mathbb{E}_p(|A_4(r,t)|) \to 0$ quand $N \to \infty$. Posons $\widehat{\tilde{e}}_k(t) = \tilde{Y}_k(t) - \widehat{Y}_{k,a}(t) = \mathbf{x}_k'(\tilde{\boldsymbol{\beta}}(t) - \hat{\boldsymbol{\beta}}_a(t))$. Avec le Lemme 2.4 et les hypothèses **B1-B5**, nous avons

$$
\begin{aligned}
\mathbb{E}_p(|A_4(r,t)|) &\leq n\mathbb{E}_p\left(\frac{1}{N^2}\sum_k\sum_l\frac{|\Delta_{kl}|}{\pi_k\pi_l}\frac{1}{\pi_{kl}}\|\mathbf{x}_k\|\|\mathbf{x}_l\|\|\tilde{\boldsymbol{\beta}}(t)-\widehat{\boldsymbol{\beta}}_a(t)\|\|\tilde{\boldsymbol{\beta}}(r)-\widehat{\boldsymbol{\beta}}_a(r)\|\right) \\
&\leq \frac{1}{n}\left[\frac{n}{\lambda^2 N}+\frac{n\max_{k\neq l}|\Delta_{kl}|}{\lambda^2\lambda_*}\right]C_4\zeta_1
\end{aligned}
$$

D'où $\mathbb{E}_p(|A_4(r,t)|) \to 0$ quand $N \to \infty$.

De la même manière, nous pouvons borner $\mathbb{E}_p(|A_2(r,t)|)$ comme suit,

$$
\begin{aligned}
\mathbb{E}_p(|A_2(r,t)|) &\leq \frac{n}{N^2}\sum_k\sum_l\frac{|\Delta_{kl}|}{\pi_k\pi_l}\frac{1}{\pi_{kl}}\mathbb{E}_p|\tilde{e}_k(t)\widehat{\tilde{e}}_l(r)| \\
&\leq \frac{n}{N^2}\sum_k\sum_l\frac{|\Delta_{kl}|}{\pi_k\pi_l}\frac{\|\mathbf{x}_l\|}{\pi_{kl}}|Y_k(t)-\tilde{Y}_k(t)|\cdot\mathbb{E}_p(\|\tilde{\boldsymbol{\beta}}(r)-\widehat{\boldsymbol{\beta}}_a(r)\|) \\
&\leq \left(\frac{\sqrt{n}}{\lambda^2 N}+\frac{\sqrt{n}\max_{k\neq l}|\Delta_{kl}|}{\lambda^2\lambda^*}\right)C_4^{1/2}\zeta_1^{1/2}\frac{1}{N}\sum_k|Y_k(t)-\tilde{Y}_k(t)|.
\end{aligned}
$$

Ainsi, il existe une constante C telle que,

$$\mathbb{E}_p(|A_2(r,t)|) \quad \leq \quad \frac{C}{\sqrt{n}}$$

et $\mathbb{E}_p(|A_2(r,t)|) \to 0$ quand $N \to \infty$. Nous pouvons montrer de la même façon que $\mathbb{E}_p(|A_3(r,t)|) \to 0$ quand $N \to \infty$.

Finalement, nous avons que pour tout $(r,t) \in [0,T]^2$,

$$n\mathbb{E}_p\left\{|\,\widehat{\gamma}_{\mathrm{MA},a}(r,t) - \gamma_{\mathrm{MA}}(r,t)\,|\right\} \to 0, \qquad \text{quand } N \to \infty. \tag{2.48}$$

D'où

$$n\mathbb{E}_p\left\{|\,\widehat{\gamma}_{\mathrm{MA},d}(r,t) - \gamma_{\mathrm{MA}}(r,t)\,|\right\} \to 0, \qquad \text{quand } N \to \infty. \tag{2.49}$$

Etape 2. Convergence uniforme de l'estimateur de la variance

La convergence ponctuelle de la fonction de variance démontrée dans l'étape précédente implique clairement la convergence en probabilité de toutes combinaisons linéaires finies : pour tout $p \in \{1,2,...\}$, pour tout $(c_1,...,c_p) \in \mathbb{R}^p$ et pour tout $(t_1,...,t_p) \in [0,T]^p$, nous avons

$$\sum_{\ell=1}^{p} c_\ell \; n\left(\widehat{\gamma}_{\mathrm{MA},a}(t_\ell,t_\ell) - \gamma_{\mathrm{MA}}(t_\ell,t_\ell)\right) \quad \to \quad 0 \tag{2.50}$$

en probabilité quand N tend à l'infini. D'après le Lemme de Cramer-Wold, le vecteur $n\left(\widehat{\gamma}_{\mathrm{MA},a}(t_1,t_1) - \gamma_{\mathrm{MA}}(t_1,t_1), \ldots, \widehat{\gamma}_{\mathrm{MA},a}(t_p,t_p) - \gamma_{\mathrm{MA}}(t_p,t_p)\right)$ converge en distribution vers 0 (dans $\mathbb{R}^p$).

Nous devons maintenant prouver que la suite de fonctions aléatoires $(n(\widehat{\gamma}_{\mathrm{MA},a}(t,t) - \gamma_{\mathrm{MA}}(t,t)))_N$ est tendue dans $C[0,T]$ en bornant les accroissements. Nous introduisons la distance suivante,

$$d_\gamma^2(t,r) \quad = \quad n^2\mathbb{E}_p(|\widehat{\gamma}_{\mathrm{MA},a}(t,t) - \gamma_{\mathrm{MA}}(t,t) - \widehat{\gamma}_{\mathrm{MA},a}(r,r) + \gamma_{\mathrm{MA}}(r,r)|^2).$$

Pour conclure, nous montrons dans ce qui suit que $d_\gamma^2(t,r) \leq C|t-r|^{2\beta}$ pour une certaine constante C et tout $(r,t) \in [0,T]^2$. A partir de (2.46), nous pouvons décomposer la distance en 4 parties.

Posons $\phi_{k,l}(t,r) = \tilde{e}_k(t)\tilde{e}_l(t) - \tilde{e}_k(r)\tilde{e}_l(r)$ et considérons $d_{A_1}^2 = \mathbb{E}_p(|A_1(t,t) - A_1(r,r)|^2)$. Nous avons

$$\begin{aligned}
d_{A_1}^2 \quad = \quad & \mathbb{E}_p\left[\frac{n^2}{N^4}\sum_k\sum_{k'}\frac{1-\pi_k}{\pi_k}\left(\frac{\mathbb{1}_k}{\pi_k}-1\right)\frac{1-\pi_{k'}}{\pi_{k'}}\left(\frac{\mathbb{1}_{k'}}{\pi_{k'}}-1\right)\phi_{k,k}(t,r)\phi_{k',k'}(t,r)\right] \\
& +2\mathbb{E}_p\left[\frac{n^2}{N^4}\sum_k\sum_{k',l':k'\neq l'}\frac{1-\pi_k}{\pi_k}\left(\frac{\mathbb{1}_k}{\pi_k}-1\right)\frac{\Delta_{k'l'}}{\pi_{k'}\pi_{l'}}\left(\frac{\mathbb{1}_{k'l'}}{\pi_{k'l'}}-1\right)\phi_{k,k}(t,r)\phi_{k',l'}(t,r)\right] \\
& +\mathbb{E}_p\left[\frac{n^2}{N^4}\sum_{k,l:k\neq l}\sum_{k',l':k'\neq l'}\frac{\Delta_{kl}}{\pi_k\pi_l}\left(\frac{\mathbb{1}_{kl}}{\pi_{kl}}-1\right)\frac{\Delta_{k'l'}}{\pi_{k'}\pi_{l'}}\left(\frac{\mathbb{1}_{k'l'}}{\pi_{k'l'}}-1\right)\phi_{k,l}(t,r)\phi_{k',l'}(t,r)\right] \\
:= \quad & b_1 + 2b_2 + b_3
\end{aligned}$$

En appliquant le Lemme 2.8, nous obtenons

$$\begin{aligned} b_1 &\leq \left(\frac{n^2}{N^3}\frac{1}{\lambda^3} + \frac{n^2}{N^2}\frac{\max_{k\neq k'}|\Delta_{kk'}|}{\lambda^4}\right)\frac{1}{N}\sum_k |\phi_{k,k}(t,r)|^2 \\ &\leq C|t-r|^{2\beta} \end{aligned} \tag{2.51}$$

et

$$\begin{aligned} b_3 &\leq \frac{(n\max_{k\neq l}|\Delta_{kl}|)^2}{\lambda^4\lambda_*^2}\max_{(k,l,k',l')\in D_{4,n}}|\mathbb{E}_p\{(\mathbb{1}_{kl}-\pi_{kl})(\mathbb{1}_{k'l'}-\pi_{k'l'})\}|\left(\frac{1}{N^2}\sum_{k,l}|\phi_{k,l}(t,r)|\right)^2 \\ &\quad + \frac{C|t-r|^{2\beta}}{N} \\ &\leq C|t-r|^{2\beta}. \end{aligned} \tag{2.52}$$

L'inégalité de Cauchy-Schwarz appliquée à b_2 avec les majorations de b_1 et b_3 nous permet d'obtenir $b_2 \leq C|t-r|^{2\beta}$. Ainsi

$$d_{A_1}^2 \leq C|t-r|^{2\beta}. \tag{2.53}$$

Nous allons maintenant majorer $d_{A_2}^2 = d_{A_3}^2 = \mathbb{E}_p(|A_2(t,t) - A_2(r,r)|^2)$. Pour cela nous définissons $\tilde{\phi}_{k,l}(t,r) = \tilde{e}_k(t)\widehat{\tilde{e}}_l(t) - \tilde{e}_k(r)\widehat{\tilde{e}}_l(r)$. En appliquant le Lemme 2.9, nous obtenons

$$\begin{aligned} d_{A_2}^2 &\leq \frac{2n^2}{N^2\lambda^4}\mathbb{E}_p\left(\frac{1}{N}\sum_k \tilde{\phi}_{k,k}(t,r)\right)^2 + \frac{2n^2\max_{k\neq l}|\Delta_{kl}|^2}{\lambda^4\lambda_*^2}\mathbb{E}_p\left(\frac{1}{N^2}\sum_{k,l}|\tilde{\phi}_{k,l}(t,r)|\right)^2 \\ &\leq C|t-r|^{2\beta}. \end{aligned} \tag{2.54}$$

Etudions le dernier terme, $d_{A_4}^2 = \mathbb{E}_p(|A_4(t,t) - A_4(r,r)|^2)$ et définissons $\widehat{\tilde{\phi}}_{k,l}(t,r) = \widehat{\tilde{e}}_k(t)\widehat{\tilde{e}}_l(t) - \widehat{\tilde{e}}_k(r)\widehat{\tilde{e}}_l(r)$. En appliquant le Lemme 2.7, nous obtenons

$$\begin{aligned} d_{A_4}^2 &\leq \frac{2n^2}{N^2\lambda^4}\mathbb{E}_p\left(\frac{1}{N}\sum_k \widehat{\tilde{\phi}}_{k,k}(t,r)\right)^2 + \frac{2n^2\max_{k\neq l}|\Delta_{kl}|^2}{\lambda^4\lambda_*^2}\mathbb{E}_p\left(\frac{1}{N^2}\sum_{k,l}|\widehat{\tilde{\phi}}_{k,l}(t,r)|\right)^2 \\ &\leq C|t-r|^{2\beta}. \end{aligned} \tag{2.55}$$

Finalement, nous pouvons déduire des inégalités (2.46), (2.53), (2.54) et (2.55), que

$$\begin{aligned} d_\gamma^2(t,r) &= n^2\mathbb{E}_p(|\widehat{\gamma}_{\mathrm{MA},a}(t,t) - \gamma_{\mathrm{MA}}(t,t) - \widehat{\gamma}_{\mathrm{MA},a}(r,r) + \gamma_{\mathrm{MA}}(r,r)|^2) \\ &\leq C|t-r|^{2\beta}. \end{aligned} \tag{2.56}$$

La fin de la preuve est une application directe du Théorème 12.3 de Billingsley (1968). Puisque $\beta > 1/2$, la séquence $(n(\widehat{\gamma}_{\mathrm{MA},a}(t,t) - \gamma_{\mathrm{MA}}(t,t)))_N$ est tendue dans $C([0,T])$.

Le théorème 8.1 de Billingsley (1968) permet de déduire que $n(\widehat{\gamma}_{\mathrm{MA},a}(t,t)-\gamma_{\mathrm{MA}}(t,t))$ converge en distribution vers 0 dans $C[0,T]$. Etant donné que $n|\widehat{\gamma}_{\mathrm{MA},d}(r,t)-\widehat{\gamma}_{\mathrm{MA},a}(r,t)| = o(1)$, nous pouvons également déduire que $n(\widehat{\gamma}_{\mathrm{MA},d}(t,t)-\gamma_{\mathrm{MA}}(t,t))$ converge en distribution vers 0 dans $C[0,T]$ (cf. Théorème 4.1 de Billingsley (1968)).

Compte tenu du fait que la fonctionnelle $\phi(f) = \sup_t |f(t)|$ est continue et bornée dans $C[0,T]$, nous appliquons la définition de la convergence en distribution pour obtenir le résultat souhaité, ainsi,

$$n\ \mathbb{E}_p\left\{\sup_{t\in[0,T]}\left|\widehat{\gamma}_{\mathrm{MA},d}(t,t)-\gamma_{\mathrm{MA}}(t,t)\right|\right\} \quad\to\quad 0.$$

□

2.4.4 Preuve de la normalité asymptotique

Les étapes de la preuve de la Proposition 2.4 sont similaires à celles de la preuve de la Proposition 2.3. Nous examinons d'abord les combinaisons finies puis nous utilisons le Lemme de Cramer-Wold. Ensuite nous montrons que la suite est tendue à partir d'une borne sur ces accroissements.

Comme nous l'avons vu dans (2.30), l'erreur d'interpolation est négligeable sous les hypothèses du schéma de discrétisation.

Ensuite, en lien avec (2.12), le Lemme 2.2 et le Lemme 2.4, nous avons clairement que, pour chaque valeur de t,

$$\sqrt{n}\,(\widehat{\mu}_{\mathrm{MA},a}(t)-\widetilde{\mu}(t)) \quad = \quad o_p(1),$$

et par conséquent, quand n tend à l'infini,

$$\sqrt{n}\left(\widehat{\mu}_{\mathrm{MA},d}(t)-\mu(t)\right) \quad\to\quad \mathcal{N}(0,\gamma_Z(t,t)) \quad \text{en distribution,}$$

où la fonction de covariance de $\widetilde{\mu}$, définie dans (2.15), satisfait $\lim_{N\to\infty} n\gamma_{\mathrm{MA}} = \gamma_Z$.

Si nous considérons maintenant p instants de discrétisation distincts $0 \le t_1 < t_2 \ldots < t_p \le 1$, nous avons immédiatement que pour tout vecteur $\mathbf{c} \in \mathbb{R}^p$, $\sqrt{n}\left(\sum_{j=1}^p c_j(\widetilde{\mu}(t_j)-\mu(t_j))\right) \to \mathcal{N}(0,\sigma_c^2)$ où

$$\sigma_c^2 = \sum_{j=1}^p\sum_{\ell=1}^p c_j c_\ell \gamma_Z(t_j,t_\ell).$$

En effet, par linéarité, il existe un vecteur de poids aléatoires $(w_1,\ldots,w_N)$ qui ne dépend pas de t tel que

$$\widetilde{\mu}(t) \quad = \quad \sum_{k\in U} w_k Y_k(t),$$

et $\sum_{j=1}^p c_j\widetilde{\mu}(t_j) = \sum_{k\in U} w_k\left(\sum_{j=1}^p c_j Y_k(t_j)\right)$ satisfait un théorème central limite, de variance asymptotique σ_c^2, sous la condition de moment **B7**. Ainsi, toute combinaison linéaire finie est asymptotiquement gaussienne et nous pouvons conclure que le vecteur

$\sqrt{n}\,(\widetilde{\mu}(t_1)-\mu(t_1),\ldots,\widetilde{\mu}(t_p)-\mu(t_p))$ est asymptotiquement gaussien en appliquant le Lemme de Cramer-Wold.

Il reste à montrer que le processus fonctionnel est tendu. Ceci est une conséquence directe de (2.38). En effet, soit $Z_n(t)=\sqrt{n}\,(\tilde{\mu}(t)-\mu(t))$, il existe une constante C telle que, pour tout $(r,t)\in[0,T]^2$,

$$\mathbb{E}_p\left([Z_n(t)-Z_n(r)]^2\right)\leq C\,|t-r|^{2\beta},$$

et, puisque $\beta>1/2$, la séquence Z_n est tendue dans $C[0,T]$ (cf. Théorème 12.3 de Billingsley (1968)) et, quand n tend à l'infini,

$$\sqrt{n}\,\{\tilde{\mu}-\mu\}\to Z \text{ en distribution dans } C[0,T] \tag{2.57}$$

où Z est un processus gaussien à valeur dans $C[0,T]$ de moyenne 0 et de fonction de covariance $\gamma_Z(r,t)=\lim_{n\to+\infty}n\gamma_{\mathrm{MA}}(r,t)$.

Etant donné que $\sqrt{n}\left(\widehat{\mu}_{\mathrm{MA},d}(t)-\mu(t)\right)=\sqrt{n}\,(\widetilde{\mu}(t)-\mu(t))+o_p(1)$, nous pouvons déduire que

$$\sqrt{n}\left\{\widehat{\mu}_{\mathrm{MA},d}-\mu\right\}\to Z \text{ en distribution dans } C[0,T] \tag{2.58}$$

quand n tend à l'infini (cf. Théorème 4.1 de Billingsley (1968))

□

Chapitre 3

Estimation de la variance de l'estimateur de la courbe moyenne de données fonctionnelles pour des plans à probabilités inégales à forte entropie

Le calcul de la variance de l'estimateur de Horvitz-Thompson pour les plans à probabilités inégales peut être très difficile car la formule d'estimation de la variance (cf. équation (1.47)) fait intervenir les probabilités d'ordre deux qui ne sont pas toujours connues. Dans le cas d'un tirage réjectif, Hájek (1964) a proposé une formule d'approximation qui ne fait intervenir que les probabilités d'inclusion d'ordre un et qui est facile à calculer. Hájek (1964) et Chen *et al.* (1994) ont montré que, lorsque les probabilités d'ordre un sont fixées, le tirage réjectif est le plan de sondage à taille fixe qui a la plus forte entropie. Matei et Tillé (2005) et Deville et Tillé (2005) ont proposé des variantes pour les plans à forte entropie. Deville et Tillé (2005) et Fuller (2009b) ont utilisé des algorithmes d'échantillonnage équilibré, ou approximativement équilibré, sur les probabilités d'inclusion pour obtenir un plan de taille fixe à probabilités d'inclusions fixées. Des arguments théoriques reliant le plan équilibré au plan réjectif sont donnés dans Deville et Tillé (2005). Dans la suite de ce chapitre, nous proposons d'étendre la formule d'approximation proposée par Hájek (1964) au cas où nous estimons une courbe moyenne à l'aide d'un plan à forte entropie.

Le chapitre est organisé de la façon suivante. Dans la section 3.1, nous présentons, pour les plans à forte entropie, l'estimateur de la covariance de l'estimateur de Horvitz-Thompson. Dans la section 3.2, nous montrons que, sous certaines hypothèses sur les probabilités d'inclusion et sur la régularité des trajectoires, l'estimateur de la fonction de variance converge uniformément vers la fonction de variance de Horvitz-Thompson.

Dans le cas d'un tirage réjectif, nous donnerons également sa vitesse de convergence. Dans la section 3.3, nous appliquerons cette méthode d'estimation sur un jeu de données de courbes de consommation. Les démonstrations seront données dans la section 3.4.

3.1 Estimation de la covariance pour les plans à fortes entropies

Considérons, comme précédemment, une population finie $U = \{1, ..., N\}$ de taille N et supposons que, pour chaque élément k de la population U, nous pouvons observer la courbe déterministe $Y_k = (Y_k(t))_{t\in[0,T]}$. Soit s un échantillon de taille fixée n, choisi aléatoirement dans U selon un plan de sondage $p(\cdot)$. Soient $\pi_k = \Pr(k \in s) > 0$ et $\pi_{kl} = \Pr(k\&l \in s)$ les probabilités d'inclusion d'ordre un et respectivement deux.

Nous supposons que la courbe moyenne μ (cf. équation (1.37)) est estimée par l'estimateur de Horvitz-Thompson fonctionnel (cf. équation (1.38)).

La formule de la covariance de l'estimateur de Horvitz-Thompson (cf. équation (1.44)) indique clairement que si, pour tout $t \in [0, T]$ et $k \in U$, la probabilité d'inclusion du premier ordre est approximativement proportionnelle à $Y_k(t)$, la variance de l'estimateur $\hat{\mu}$ sera faible. Par conséquent, si nous disposons d'une variable auxiliaire X supposée à valeur positive et connue pour l'ensemble des individus de la population et si X est corrélée à notre variable d'intérêt, il est intéressant de considérer un plan de sondage où les probabilités d'inclusion sont données par

$$\pi_k = n\frac{x_k}{\sum_U x_k}, \quad \forall k \in U. \tag{3.1}$$

Nous avons vu dans le chapitre précédent différents plans de sondage qui permettent de tirer un échantillon de taille fixe n et tels que les probabilités d'inclusion vérifient l'équation (3.1). Dans la suite de ce chapitre, nous allons nous intéresser aux plans à forte entropie (cf. Définition 1.1). Chen *et al.* (1994) ont prouvé que lorsque les probabilités d'inclusion d'ordre un sont fixées, le tirage réjectif est le plan à taille fixe qui a la plus forte entropie. Ce résultat clé est lié à l'approximation uniforme des probabilités d'inclusion d'ordre deux

$$\pi_{kl} = \pi_k\pi_l\left\{1 - \frac{(1-\pi_k)(1-\pi_l)}{d(\pi)}[1 + o(1)]\right\}, \tag{3.2}$$

où $d(\pi) = \sum_U \pi_k(1-\pi_k)$. Cette approximation est satisfaite pour le tirage réjectif et le tirage de Sampford-Durbin (Hájek (1981)) et s'avère être très efficace quand la taille de l'échantillon est grande (et par conséquent quand d est grand) et que l'entropie du plan de sondage est proche de l'entropie maximale.

En introduisant l'approximation (3.2) dans l'équation (1.44), nous obtenons l'approximation de Hájek $\gamma_H(r,t)$ de la fonction de covariance $\gamma_p(r,t)$ dans le cadre fonc-

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

$$\gamma_H(r,t) = \frac{1}{N^2}\left[\sum_U \frac{Y_k(t)Y_k(r)}{\pi_k}(1-\pi_k) - \frac{1}{d(\pi)}\sum_U\sum_U (1-\pi_k)(1-\pi_l)Y_k(t)Y_l(r)\right]. \tag{3.3}$$

La variance $\gamma_H(t,t)$ est toujours positive (Bondesson *et al.* (2006)) et un estimateur de la covariance est donné par

$$\hat{\gamma}_H(r,t) = \frac{1}{N^2}\frac{\hat{d}(\pi)}{d(\pi)}\left[\sum_s \frac{1-\pi_k}{\pi_k^2}Y_k(t)Y_k(r) - \frac{1}{\hat{d}(\pi)}\sum_s\sum_s \frac{1-\pi_k}{\pi_k}\frac{1-\pi_l}{\pi_l}Y_k(t)Y_l(r)\right] \tag{3.4}$$

où $\hat{d}(\pi) = \sum_s(1-\pi_k)$. C'est l'extension, au cas fonctionnel, de l'estimateur de la variance proposé par Berger (1998a) dans le cas réel (cf. équation (1.13)).

Nous pouvons facilement montrer la propriété suivante.

Proposition 3.1. *Si, pour tout $t \in [0,T]$, il existe une constante c_t tel que $Y_k(t) = c_t\pi_k$ alors $\gamma_H(r,t) = 0$ et $\widehat{\gamma}_H(r,t) = 0$.*

Lorsque nous travaillons sur un jeu de données discrétisées, la covariance est estimée à l'aide des courbes interpolées $Y_{k,d}$ (cf. équation (1.45))

$$\hat{\gamma}_{H,d}(r,t) = \frac{1}{N^2}\frac{\hat{d}(\pi)}{d(\pi)}\left[\sum_s \frac{1-\pi_k}{\pi_k^2}Y_{k,d}(t)Y_{k,d}(r) - \frac{1}{\hat{d}(\pi)}\sum_s\sum_s \frac{1-\pi_k}{\pi_k}\frac{1-\pi_l}{\pi_l}Y_{k,d}(t)Y_{l,d}(r)\right], \tag{3.5}$$

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

3.2 Propriétés asymptotiques

3.2.1 Hypothèses

Pour démontrer les propriétés asymptotiques de l'estimateur de Horvitz-Thompson pour les plans qui vérifient l'équation (3.2), nous nous plaçons dans le cadre de super-population introduit par Isaki et Fuller (1982) (cf. section 1.5) et nous adoptons l'approche asymptotique proposée par Hájek (1964), nous supposons que $d(\pi) \to \infty$. Cette hypothèse implique que $n \to \infty$ et $N - n \to \infty$. Il est également nécessaire d'introduire les hypothèses suivantes.

A1. Nous supposons que $\lim_{N\to\infty} \frac{n}{N} = \pi \in]0,1[$.

A2. Nous supposons que $\min_{k\in U} \pi_k \geq \lambda > 0$, $\min_{k\neq l} \pi_{kl} \geq \lambda^* > 0$ et

$$\pi_{kl} = \pi_k\pi_l\left\{1 - \frac{(1-\pi_k)(1-\pi_l)}{d(\pi)}[1+o(1)]\right\}.$$

A3. Il existe deux constantes positives C_1 et C_2 et $\beta > 1/2$ telles que, pour tout N et pour tout $(r,t) \in [0,T] \times [0,T]$,

$$\frac{1}{N}\sum_{k\in U}(Y_k(0))^2 < C_1 \quad \text{et} \quad \frac{1}{N}\sum_{k\in U}(Y_k(t)-Y_k(r))^2 < C_2|t-r|^{2\beta}.$$

Les hypothèses **A1** et **A2** sont des hypothèses classiques en sondage et elles concernent les probabilités d'inclusion d'ordre un et deux. De plus, elles sont satisfaites pour certains plans de sondage à taille fixe (Hájek (1981)) et elles impliquent que $cn \leq d(\pi) \leq n$ pour une certaine constante c strictement positive. L'hypothèse **A2** implique que $\limsup_{N\to\infty} n \max_{k\neq l} |\pi_{kl} - \pi_k\pi_l| < C_1 < \infty$, et assure que $\pi_{kl} \leq \pi_k\pi_l$ ainsi l'estimateur de la variance proposé par Sen (1953) et Yates et Grundy (1953) est toujours positif.

L'hypothèse **A3** a déjà été introduite dans Cardot et Josserand (2011). Même si la convergence ponctuelle peut être prouvée sans aucune condition sur β, cette condition de régularité est nécessaire pour obtenir la convergence uniforme de l'estimateur de Horvitz-Thompson.

3.2.2 Convergence et propriétés asymptotiques

Cardot et Josserand (2011) ont montré que, sous certaines hypothèses, l'estimateur $\hat{\mu}_d$ est asymptotiquement non biaisé par rapport au plan et uniformément convergent.

Proposition 3.2. Cardot et Josserand (2011)
Si les hypothèses **A1**-**A3** *sont vérifiées et si le schéma de discrétisation satisfait* $\max_{i=1,\dots,D_N-1} |t_{i+1}-t_i|^{2\beta} = o(n^{-1})$ *alors, il existe une constante* C *telle que*

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

Le résultat ci-dessus montre que si la grille est assez fine, la vitesse de convergence paramétrique peut-être obtenue de façon uniforme. L'hypothèse ajoutée montre que la grille doit être d'autant plus dense que β est petit. De plus, la condition $\beta > 1/2$ dans l'hypothèse **A3**, qui exige que les trajectoires soient assez régulières (sans nécessairement être dérivables), n'est pas forte puisque le cas $\beta = 1/2$ correspond à des trajectoires browniennes.

On souhaite maintenant obtenir la convergence de l'estimateur $\hat{\gamma}_{H,d}$ de la fonction de covariance $\gamma_p(r,t)$. Pour cela, il est nécessaire d'introduire de nouvelles hypothèses sur les moments d'ordre 4 des trajectoires et sur les probabilités d'inclusion d'ordre supérieur,

A4. Il existe deux constantes positives C_3 et C_4 telles que, pour tout N et pour tout $(r,t) \in [0,T] \times [0,T]$,

$$\frac{1}{N}\sum_{k\in U}(Y_k(0))^4 < C_3 \quad \text{et} \quad \frac{1}{N}\sum_{k\in U}(Y_k(t)-Y_k(r))^4 < C_4|t-r|^{4\beta}.$$

A5. Nous supposons que

$$\lim_{N\to\infty} \max_{(k_1,l_1,k_2,l_2)\in D_{4,N}} |\mathbb{E}_p\left[(\mathbb{1}_{k_1l_1} - \pi_{k_1}\pi_{l_1})(\mathbb{1}_{k_2l_2} - \pi_{k_2}\pi_{l_2})\right]| \to 0$$

où $D_{t,N}$ représente l'ensemble de tous les t-tuples distincts $(i_1, ..., i_t)$ de U.

L'hypothèse **A5** est vérifiée pour le sondage aléatoire simple sans remise, le sondage stratifié et plus généralement pour les plans à forte entropie. Considérons la divergence de Kullback-Leibler $K(p, p_{rej})$,

$$K(p, p_{rej}) = \sum_s p(s) \ln\left(\frac{p(s)}{p_{rej}(s)}\right). \tag{3.6}$$

Celle-ci permet de mesurer la dissimilarité entre la distribution du plan p est celle du plan à entropie maximum c'est-à-dire le tirage réjectif p_{rej}. Nous remarquons que cette divergence n'est pas symétrique.

Proposition 3.3. *Si $d(\pi) \to \infty$, alors*

$$\max_{(k_1,l_1,k_2,l_2)\in D_{4,N}} |\mathbb{E}_p\left[(\mathbb{1}_{k_1l_1} - \pi_{k_1}\pi_{l_1})(\mathbb{1}_{k_2l_2} - \pi_{k_2}\pi_{l_2})\right]| \leq \frac{C}{d(\pi)} + \sqrt{K(p, p_{rej})/2}$$

pour une certaine constante C.

Une conséquence directe de la Proposition 3.3 est que l'hypothèse **A5** est vérifiée pour le tirage réjectif ainsi que pour le tirage de Sampford-Durbin, dont la divergence de Kullback-Leibler, par rapport au tirage réjectif, tend vers 0 quand la taille n de l'échantillon tend à l'infini (Berger (1998b)).

Proposition 3.4. *Supposons que les hypothèses* **A1**-**A5** *sont vérifiées et que le schéma de discrétisation vérifie $\max_{i=\{1,..,D_N-1\}} |t_{i+1} - t_i| = o(1)$. Quand N tend à l'infini,*

$$n\mathbb{E}_p\left\{|\,\widehat{\gamma}_{H,d}(r,t) - \gamma_p(r,t)\,|\right\} \to 0$$

et

$$n\mathbb{E}_p\left\{\sup_{t\in[0,T]} |\,\widehat{\gamma}_{H,d}(t,t) - \gamma_p(t,t)\,|\right\} \to 0.$$

L'estimateur $\widehat{\gamma}_{H,d}(r,t)$ de la covariance $\gamma_p(r,t)$ est ponctuellement convergent pour tout $(r,t) \in [0,T] \times [0,T]$ et l'estimateur fonctionnel de la variance $\widehat{\gamma}_{H,d}(t,t)$ est uniformément convergent vers $\gamma_p(t,t)$. Pour obtenir les taux de convergence, il faudrait introduire de nouvelles hypothèses. La preuve est donnée dans la section 3.4.

Pour le tirage réjectif, Boistard *et al.* (2012) ont donné une approximation des probabilités d'inclusion multiple qui nous permet d'obtenir le résultat suivant :

Proposition 3.5. *Supposons que l'échantillon s est sélectionné à l'aide d'un tirage réjectif. Supposons que les hypothèses* **A1-A4** *sont vérifiées et que le schéma de discrétisation satisfait* $\max_{i=\{1,..,D_N-1\}} |t_{i+1}-t_i|^{2\beta} = O(n^{-1})$. *Alors, pour tout* $(r,t) \in [0,T]^2$,

$$n^3 \ \mathbb{E}_p\left[\left(\widehat{\gamma}_{H,d}(r,t) - \gamma_p(r,t)\right)^2\right] \leq C$$

pour certaine constante positive C.

On peut également remarquer que l'erreur d'approximation de la variance γ_p par la formule de Hájek γ_H est asymptotiquement négligeable comparée à l'erreur d'échantillonnage (cf. preuve de la Proposition 3.5).

Pour montrer que l'estimateur de Horvitz-Thompson $\hat{\mu}_d$ satisfait un théorème centrale limite fonctionnelle Cardot et Josserand (2011) ont introduit une nouvelle hypothèse.

A6. Il existe $\delta > 0$, tel que $N^{-1}\sum_{k\in U}|Y_k(t)|^{2+\delta} < \infty$ pour tout $t \in [0,T]$, et $\{\gamma_p(t,t)\}^{-1/2}\{\hat{\mu}(t)-\mu(t)\} \to \mathcal{N}(0,1)$ en distribution quand N tend à l'infini.

Proposition 3.6. *Cardot et Josserand (2011)*
Si les hypothèses **A1-A3** *et* **A6** *sont vérifiées et si le schéma de discrétisation satisfait* $\max_{i=\{1,..,D_N-1\}} |t_{i+1}-t_i|^{2\beta} = o(n^{-1})$, *alors*

$$\sqrt{n}(\hat{\mu}_d - \mu) \to Z \text{ en distribution dans } C[0,T]$$

où Z est un processus gaussien prenant ses valeurs dans $C[0,T]$ de moyenne 0 et de fonction de covariance $\gamma_Z(r,t) = \lim_{N\to\infty} n\gamma_p(r,t)$.

Ce dernier résultat nous servira à justifier la construction de la bande de confiance par simulation de processus gaussien dans le chapitre 4.

3.3 Exemple : estimation de la variance de la courbe de consommation électrique

Lorsque l'échantillon est tiré à l'aide de l'algorithme du cube, Deville et Tillé (2005) suggèrent d'utiliser des approximations de la variance de l'estimateur de Horvitz-Thompson basées sur les idées de Hájek. Dans cette section, nous allons calculer la précision de l'approximation de Hájek $\hat{\gamma}_{H,d}$ lorsqu'on dispose d'une grande population de données fonctionnelles.

Nous disposons ici d'une population de $N = 15055$ courbes de consommation électrique mesurées toutes les demi-heures pendant une semaine, soit $d = 336$ points de discrétisation. La consommation moyenne de la semaine précédente X, connue pour chaque individu k de la population U, est utilisée comme variable auxiliaire. Cette variable est très corrélée à la consommation d'électricité mesurée pendant notre semaine

d'étude (la corrélation ponctuelle est supérieure à 0.80 cf. Figure 1.1). Notons que la consommation totale passée sera facilement mesurable et transmise après la mise en place des nouveaux compteurs.

Nous supposons que l'échantillon s de taille n est tiré à l'aide de l'algorithme du cube équilibré sur la variable $\boldsymbol{\pi} = (\pi_1, ..., \pi_N)$ avec $\pi_k = n\frac{x_k}{\sum_U x_k}$. Afin d'obtenir un plan de sondage proche de l'entropie maximale, un tri aléatoire de la population est effectué avant de tirer l'échantillon s (cf. Tillé (2011)).

La courbe moyenne de consommation est estimée à l'aide de l'estimateur $\hat{\mu}_d$ défini par l'équation (1.46). Les probabilités d'inclusion d'ordre deux étant inconnues, une estimation empirique de la covariance γ_p est obtenue à partir de $J = 10000$ simulations

$$\gamma_{emp}(r,t) = \frac{1}{J-1}\sum_{j=1}^{J}(\hat{\mu}_{d,j}(t) - \hat{\bar{\mu}}_d(t))(\hat{\mu}_{d,j}(r) - \hat{\bar{\mu}}_d(r)) \tag{3.7}$$

avec $\hat{\mu}_{d,j}(t) = \frac{1}{N}\sum_{k\in s_j}\frac{Y_{k,d}(t)}{\pi_k}$, $\hat{\bar{\mu}}_d(t) = \frac{1}{J}\sum_{j=1}^{J}\hat{\mu}_{d,j}(t)$ et $(r,t) \in [0,T]$.

Afin d'évaluer la performance de l'estimateur $\hat{\gamma}_{H,d}(t,t)$ de la fonction de variance γ_p, nous considérons différentes tailles d'échantillon $n = 250$ ($d(\pi) = 241.2$), $n = 500$ ($d(\pi) = 464.7$) et $n = 1500$ ($d(\pi) = 1202.3$). Pour chaque taille n d'échantillon, nous tirons $I = 10000$ échantillons. La précision de l'estimateur $\hat{\gamma}_{H,d}$ est évaluée à l'aide du critère quadratique suivant :

$$\begin{aligned} R_2(\hat{\gamma}_{H,d}) &= \frac{1}{336}\sum_{i=1}^{336}\frac{|\hat{\gamma}_{H,d}(t_i,t_i) - \gamma_{emp}(t_i,t_i)|^2}{\gamma_{emp}(t_i,t_i)^2} \\ &\simeq \int \frac{|\hat{\gamma}_{H,d}(t,t) - \gamma_{emp}(t,t)|^2}{\gamma_{emp}(t,t)^2}dt \end{aligned}$$

Nous considérons ensuite l'erreur quadratique moyenne relative

$$\begin{aligned} RMSE(\hat{\gamma}_{H,d}) &= \frac{1}{I-1}\sum_{i=1}^{I}R_{2,i}(\hat{\gamma}_{H,d}) \\ &= BR(\hat{\gamma}_{H,d})^2 + VR(\hat{\gamma}_{H,d}) \end{aligned} \tag{3.8}$$

où

$$BR(\hat{\gamma}_{H,d})^2 = \frac{1}{336}\sum_{i=1}^{336}\left(\frac{\bar{\hat{\gamma}}_{H,d}(t,t) - \gamma_{emp}(t_i,t_i)}{\gamma_{emp}(t_i,t_i)}\right)^2$$

est le biais relatif de l'estimateur $\hat{\gamma}_{H,d}$, $VR(\hat{\gamma}_{H,d})$, sa variance relative et $\bar{\hat{\gamma}}_{H,d}(t,t) = \frac{1}{I}\sum_{i=1}^{I}\hat{\gamma}_{H,d,i}(t,t)$.

Les erreurs d'estimation sont présentées dans le tableau 3.1 pour les 3 tailles d'échantillons. L'erreur médiane est faible et décroît logiquement avec la taille n de l'échantillon. Cependant l'erreur moyenne est relativement élevée. Les quantiles à 95% indiquent clairement que pour un très petit nombre d'échantillons l'erreur d'estimation de la variance est très importante. On remarque également que le biais relatif

Taille de l'échantillon	$RMSE(\hat{\gamma}_{H,d})$	$BR(\hat{\gamma}_{H,d})^2$	$R_2(\hat{\gamma}_{H,d}(t,t))$				
			Q5	Q25	médiane	Q75	Q95
250	0.9473	0.0004	0.0188	0.0298	0.0446	0.0748	0.4326
500	0.3428	0.0002	0.0121	0.0191	0.0278	0.0456	0.3510
1500	0.1406	0.0003	0.006	0.0097	0.0144	0.0272	0.0929

TABLE 3.1 – $RMSE(\hat{\gamma}_{H,d})$, $RB(\hat{\gamma}_{H,d})^2$ et l'erreur d'estimation $R_2(\hat{\gamma}_{H,d})$ pour différentes tailles d'échantillon, avec $I = 10000$ échantillons.

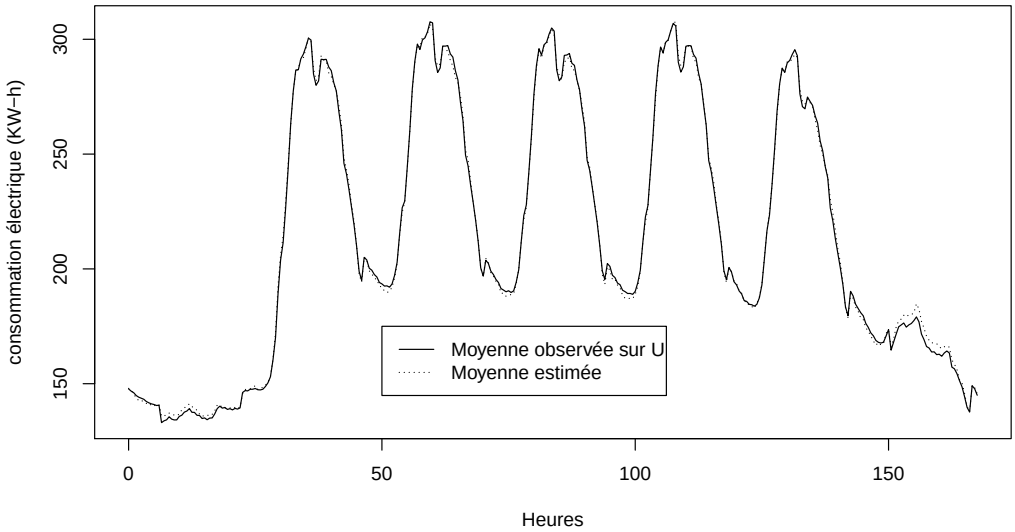


FIGURE 3.1 – Courbe moyenne de la consommation sur la population et son estimation obtenue à l'aide de l'échantillon s' de taille $n = 1500$

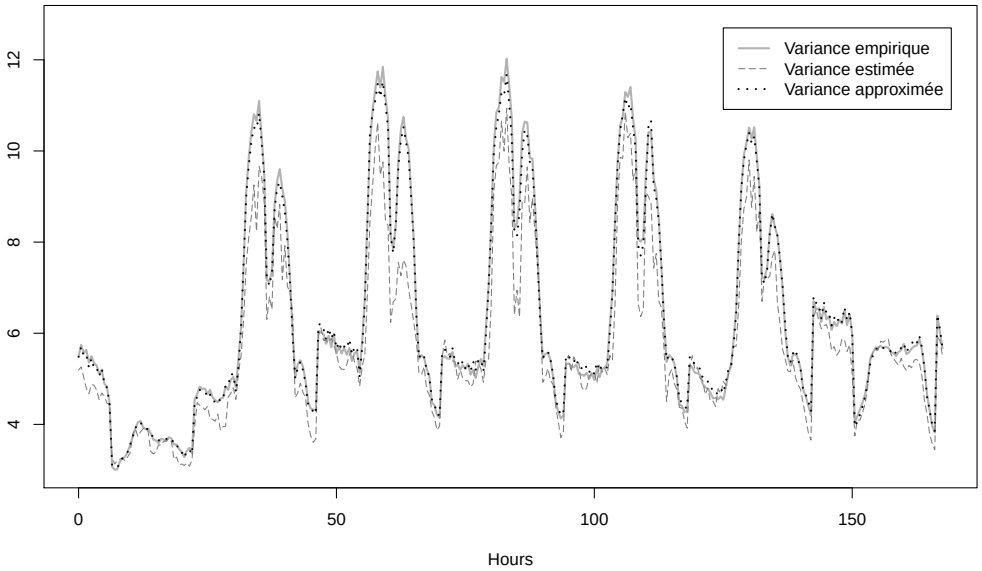


FIGURE 3.2 – Variance empirique γ_{emp}, approximation de Hájek γ_H et la variance estimée $\hat{\gamma}_{H,d}$ obtenue à l'aide de l'échantillon s' de taille $n = 1500$

est négligeable par rapport à la variance de l'estimateur (ce qui est en accord avec la Proposition 3.5).

Sur la Figure 3.1, nous avons représenté la courbe moyenne de consommation observée sur la population U ainsi qu'une estimation obtenue à l'aide d'un échantillon, noté s', de taille n=1500 pour lequel l'erreur $R_2(\hat{\gamma}_{H,d})$ est proche de la valeur médiane. Sur la Figure 3.2, nous avons représenté la fonction de variance empirique γ_{emp} de l'estimateur $\hat{\mu}_d$, son approximation d'Hájek γ_H ainsi que son estimation $\hat{\gamma}_{H,d}$ obtenue à partir de l'échantillon s'. Sur la Figure 3.3, nous avons représenté l'erreur d'estimation $\gamma_{emp}(t,r) - \hat{\gamma}_{H,d}(t,r)$ commise lorsqu'on estime la courbe moyenne à partir de l'échantillon s' pour $(t,r) \in [0,T]\text{x}[0,T]$. Sur la Figure 3.4, nous avons représenté l'erreur d'approximation $\gamma_{emp}(t,r) - \gamma_H(t,r)$ commise lorsque les échantillons sont de taille $n = 1500$ pour $(t,r) \in [0,T]\text{x}[0,T]$. L'erreur d'approximation est très faible comparée à l'erreur d'échantillonnage (cf. Figures 3.2 et 3.4). Nous sous-estimons la fonction de covariance aux instants où nous avons les pics de consommation et nous la surestimons légèrement aux autres instants (cf. Figure 3.3).

L'erreur $R_2(\hat{\gamma}_{H,d})$ médiane est très faible (de l'ordre de 1.5% pour $n = 1500$) cependant nous constatons qu'en moyenne elle reste très élevée (14% pour $n = 1500$) malgré l'augmentation de la taille de l'échantillon. Cela est dû à la présence dans notre population d'étude d'individus qui ont à la fois une faible probabilité d'inclusion π_k (cf. Figure 3.5) et une consommation Y_k ponctuellement très élevée et inversement, d'individus qui ont une forte probabilité d'inclusion π_k et une faible consommation. Par exemple, un individu qui revient de vacances durant notre semaine d'étude aura

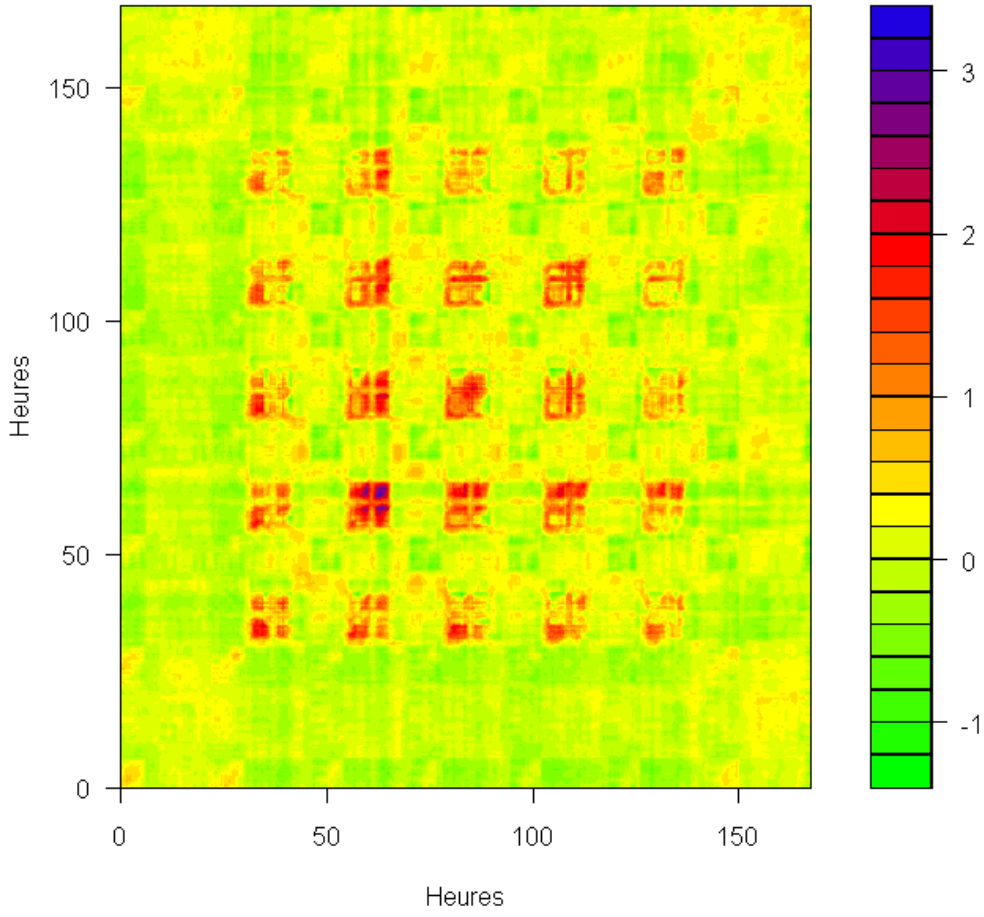


FIGURE 3.3 – Erreur d'estimation $\gamma_{emp}(t,r) - \hat{\gamma}_{H,d}(t,r)$ obtenue à l'aide de l'échantillon s' de taille $n = 1500$

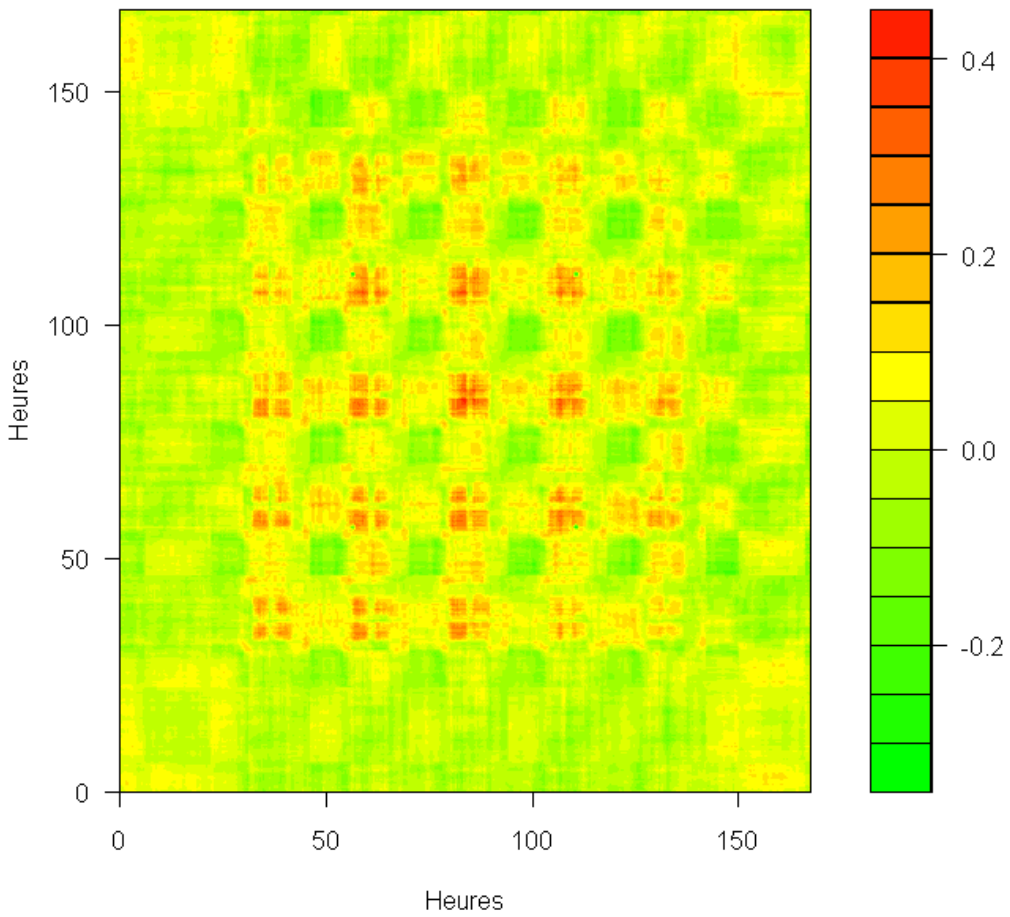


FIGURE 3.4 – Erreur d'approximation $\gamma_{emp}(t,r) - \gamma_H(t,r)$ pour $n = 1500$

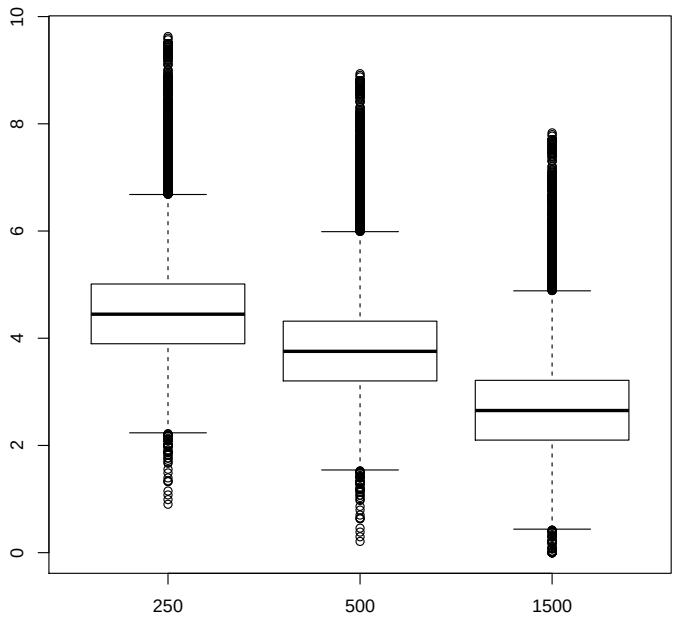


FIGURE 3.5 – Représentation du logarithme des poids de sondage ($\log(1/\pi_k)$) pour les différentes tailles d'échantillon

une valeur faible de π_k et un niveau moyen de consommation très différent de celui de la semaine précédente. La présence de tels individus dans l'échantillon conduit alors à une mauvaise estimation de la courbe moyenne et cette erreur est de plus "amplifiée" lors de l'estimation de la variance.

En augmentant la taille de l'échantillon, nous avons diminué le poids de ces individus (cf. Figure 3.5) et par conséquent nous avons diminué la variance de l'estimateur $\hat{\gamma}_{H,d}$ mais cela n'est pas suffisant. Notre plan de sondage et notre estimateur ne sont pas robustes. On pourrait sans doute obtenir des estimateurs de la variance plus stables en utilisant des méthodes permettant de corriger le poids de sondage des individus influents (cf. par exemple Beaumont et Rivest (2009)).

3.4 Preuves

Tout au long des preuves, nous utiliserons la lettre C pour désigner une constante générique dont la valeur peut varier d'un endroit à l'autre. Nous définissons $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ et $\Delta_{kk} = \pi_k(1-\pi_k)$.

3.4.1 Quelques lemmes utiles

Lemme 3.1. *Si l'hypothèse* **A4** *est vérifiée alors il existe une constante* ζ_1 *telle que*

$$\frac{1}{N}\sum_{k\in U}|\phi_{k,k}(t,r)|^2 \leq \zeta_1|t-r|^{2\beta}$$

où $\phi_{k,k}(t,r) = Y_k(t)Y_k(t) - Y_k(r)Y_k(r)$.

Démonstration. En utilisant l'inégalité de Cauchy-Schwarz, nous obtenons

$$\begin{aligned}\frac{1}{N}\sum_{k\in U}|\phi_{k,k}(t,r)|^2 &\leq \frac{2}{N}\left\{\sum_U |Y_k(t)-Y_k(r)|^2|Y_k(t)|^2 + \sum_U |Y_k(t)-Y_k(r)|^2|Y_k(r)|^2\right\}\\ &\leq 2\left[\left(\frac{1}{N}\sum_U |Y_k(t)-Y_k(r)|^4\right)^{1/2}\left(\frac{1}{N}\sum_U |Y_k(t)|^4\right)^{1/2}\right.\\ &\left.+\left(\frac{1}{N}\sum_U |Y_k(t)-Y_k(r)|^4\right)^{1/2}\left(\frac{1}{N}\sum_U |Y_k(r)|^4\right)^{1/2}\right].\end{aligned}$$

Sous l'hypothèse **A4**, nous obtenons que, pour une certaine constante ζ_1,

$$\frac{1}{N}\sum_{k\in U}|\phi_{k,k}(t,r)|^2 \leq \zeta_1|t-r|^{2\beta}.$$

□

Lemme 3.2. *Si l'hypothèse **A3** est vérifiée alors il existe une constante* ζ_2 *telle que*

$$\left(\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}|\phi_{k,l}(t,r)|\right)^2 \leq \zeta_2|t-r|^{2\beta}$$

où $\phi_{kl}(t,r) = Y_k(t)Y_l(t) - Y_k(r)Y_l(r)$.

Démonstration. La démonstration est similaire à celle du Lemme 3.1 et sera donc omise. □

Lemme 3.3. *Supposons que les hypothèses **A1** et **A2** sont vérifiées.*

$$\mathbb{E}_p((\hat{d}(\pi)-d(\pi))^2) \leq \left(\frac{1}{\lambda} + \frac{\max_{k\neq l}|\Delta_{kl}|}{\lambda^2}d(\pi)\right)d(\pi).$$

Démonstration. Sous les hypothèses **A1** et **A2**,

$$\begin{aligned}\mathbb{E}_p((\hat{d}(\pi)-d(\pi))^2) &= \sum_U\sum_U \frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}\pi_k(1-\pi_k)\pi_l(1-\pi_l)\\ &\leq \frac{1}{\lambda}\sum_U \pi_k^2(1-\pi_k)^2 + \frac{\max_{k\neq l}|\Delta_{kl}|}{\lambda^2}\left(\sum_U \pi_k(1-\pi_k)\right)^2\\ &\leq \left(\frac{1}{\lambda} + \frac{\max_{k\neq l}|\Delta_{kl}|}{\lambda^2}d(\pi)\right)d(\pi).\end{aligned}$$

□

3.4.2 Preuve de la proposition 3.3

Considérons dans un premier temps le cas du tirage réjectif $p_{rej}(s)$ et montrons que l'hypothèse **A5** est vérifiée si $d(\pi)$ tend vers l'infini. D'après le théorème 1 de Boistard *et al.* (2012) et l'hypothèse **A2**, nous avons

$$\mathbb{E}_p\left(\mathbb{1}_{k_1k_2l_1l_2}\right) - \pi_{k_1}\pi_{k_2}\pi_{l_1}\pi_{l_2} = O(d(\pi)^{-1})$$

uniformément pour $(k_1, l_1, k_2, l_2) \in D_{4,N}$.

Comme $\pi_{k_1}\pi_{k_2} - \pi_{k_1k_2} = O(d(\pi)^{-1})$ et $\pi_{l_1}\pi_{l_2} - \pi_{l_1l_2} = O(d(\pi)^{-1})$ uniformément pour $(k_1, l_1, k_2, l_2) \in D_{4,N}$, nous obtenons directement que, pour le tirage réjectif,

$$\max_{(k_1,l_1,k_2,l_2)\in D_{4,N}} |\mathbb{E}_p\left[(\mathbb{1}_{k_1l_1} - \pi_{k_1}\pi_{l_1})(\mathbb{1}_{k_2l_2} - \pi_{k_2}\pi_{l_2})\right]| \leq \frac{C}{d(\pi)},$$

pour une certaine constante C.

Si nous considérons maintenant un autre plan de sondage $p_N(s)$, nous avons avec l'inéqualité de Pinsker (cf. Théorème 6.1 dans Kemperman (1969)) une majoration de la distance de variation totale,

$$\sup_{A\in\mathcal{A}_{\mathcal{N}}} |p_N(A) - p_{rej}(A)| \leq \sqrt{K(p_N, p_{rej})/2}$$

où $\mathcal{A}_{\mathcal{N}}$ est l'ensemble de toutes les partitions de U_N. En considérant le cas particulier $A = \{(k_1, l_1, k_2, l_2) \in D_{4,N}\}$, et en posant $\pi_{k_1k_2l_1l_2} = p_N(A)$ et $\pi^{rej}_{k_1k_2l_1l_2} = p_{rej}(A)$, nous obtenons directement que

$$\sup_{(k_1,l_1,k_2,l_2)\in D_{4,N}} \left|\pi_{k_1k_2l_1l_2} - \pi^{rej}_{k_1k_2l_1l_2}\right| \leq \sqrt{K(p_N, p_{rej})/2}.$$

La preuve est terminée.

3.4.3 Preuve de la proposition 3.4

Dans un premier temps, nous allons montrer que la variable aléatoire $n(\widehat{\gamma}_{H,d}(t,t) - \gamma_p(t,t))$ converge en distribution vers zéro dans l'espace des fonctions continues muni de la norme sup, noté $C[0,T]$. Etant donné que la fonctionnelle $\phi(f) = \sup_t |f(t)|$ est continue et bornée dans $C[0,T]$, nous déduirons de la définition de la convergence en distribution dans $C[0,T]$ le résultat annoncé.

La preuve est décomposée en 2 étapes. Nous montrons d'abord la convergence ponctuelle, en considérant la convergence de toutes les combinaisons linéaires finies, et ensuite nous montrons que la suite est tendue.

Etape 1. Convergence ponctuelle

Nous voulons montrer, que pour chaque couple $(t,r) \in [0,T]^2$, nous avons

$$n\mathbb{E}_p\left\{|\,\widehat{\gamma}_{H,d}(r,t) - \gamma_p(r,t)\,|\right\} \to 0, \quad \text{quand } N \to \infty.$$

Décomposons

$$n(\widehat{\gamma}_{\mathrm{H},d}(r,t)-\gamma_p(r,t))=n(\widehat{\gamma}_{\mathrm{H},d}(r,t)-\widehat{\gamma}_H(r,t))+n(\widehat{\gamma}_H(r,t)-\gamma_p(r,t))$$

et étudions séparément les erreurs d'interpolation et d'estimation.

Erreur d'interpolation

Supposons que $t\in[t_i,t_{i+1}[$ et $r\in[t_{i'},t_{i'+1}[$.

$$\begin{aligned}n|\hat{\gamma}_{H,d}(r,t)-\hat{\gamma}_H(r,t)|\leq{}&\frac{n}{N^2}\frac{\hat{d}(\pi)}{d(\pi)}\sum_U\frac{1-\pi_k}{\pi_k^2}|Y_{k,d}(t)Y_{k,d}(r)-Y_k(t)Y_k(r)|\\&+\frac{n}{N^2}\frac{1}{d(\pi)}\sum_U\sum_U\frac{1-\pi_k}{\pi_k}\frac{1-\pi_l}{\pi_l}|Y_{k,d}(t)Y_{l,d}(r)-Y_k(t)Y_l(r)|.\end{aligned}$$

Après avoir remarqué que $\hat{d}(\pi)\leq\frac{1}{\lambda}d(\pi)$, $\frac{1}{d(\pi)}\leq\frac{1}{cn}$ et

$$\begin{aligned}|Y_{k,d}(t)Y_{l,d}(r)-Y_k(t)Y_l(r)|\leq{}&|Y_k(t_i)-Y_k(t)||Y_l(t_{i'})|+|Y_l(t_{i'})-Y_l(r)||Y_k(t)|\\&+|Y_k(t_{i+1})-Y_k(t_i)|\,[|Y_l(t_{i'+1})|+2|Y_l(t_{i'})|]\\&+|Y_k(t_i)||Y_l(t_{i'+1})-Y_l(t_{i'})|\end{aligned}$$

nous avons, sous les hypothèses **A1-A4**,

$$\begin{aligned}n|\hat{\gamma}_{H,d}(r,t)-\hat{\gamma}_H(r,t)|&\leq\left(\frac{n}{N}\frac{1}{\lambda}+\frac{n}{d(\pi)}\right)\frac{C}{\lambda^2}[|t_i-t|^\beta+|t_{i'}-r|^\beta+|t_{i'+1}-t_{i'}|^\beta+3|t_{i+1}-t_i|^\beta]\\&\leq 2\left(\frac{n}{N}\frac{1}{\lambda^3}+\frac{1}{c}\right)\frac{C}{\lambda^2}[2|t_{i+1}-t_i|^\beta+|t_{i'+1}-t_{i'}|^\beta].\end{aligned}\tag{3.9}$$

Ainsi, sous l'hypothèse du schéma de discrétisation,

$$n|\hat{\gamma}_{H,d}(r,t)-\hat{\gamma}_H(r,t)|=o(1).$$

Considérons la décomposition suivante

$$|\,\hat{\gamma}_H(r,t)-\gamma_p(r,t)\,|\leq|\,\hat{\gamma}_H(r,t)-\gamma_H(r,t)\,|+|\,\gamma_H(r,t)-\gamma_p(r,t)\,|\tag{3.10}$$

et étudions séparément ces 2 termes d'erreurs.

Erreur d'approximation

Nous montrons d'abord que, pour chaque $(r,t)\in[0,T]^2$,

$$n\,|\,\gamma_H(r,t)-\gamma_p(r,t)\,|=o(1).$$

En introduisant l'approximation (3.2)

$$\pi_{kl}-\pi_k\pi_l=-\pi_k\pi_l\frac{(1-\pi_k)(1-\pi_l)}{d(\pi)}+\frac{c_{kl}}{d(\pi)}\tag{3.11}$$

où $\max_{k,l}|c_{kl}| \to 0$, dans la fonction de covariance (1.44), nous obtenons

$$\begin{aligned}\gamma_p(r,t) &= \frac{1}{2}\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\frac{\pi_k\pi_l(1-\pi_k)(1-\pi_l)-c_{kl}}{d(\pi)}\left(\frac{Y_k(r)}{\pi_k}-\frac{Y_l(r)}{\pi_l}\right)\left(\frac{Y_k(t)}{\pi_k}-\frac{Y_l(t)}{\pi_l}\right)\\ &= \gamma_H(r,t)-\frac{1}{2}\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\frac{c_{kl}}{d(\pi)}\left(\frac{Y_k(r)}{\pi_k}-\frac{Y_l(r)}{\pi_l}\right)\left(\frac{Y_k(t)}{\pi_k}-\frac{Y_l(t)}{\pi_l}\right)\end{aligned}$$

Finalement, nous déduisons directement des hypothèses **A1**-**A3** que

$$d(\pi)\ |\gamma_H(r,t)-\gamma_p(r,t)| = o(1). \tag{3.12}$$

et

$$n\ |\gamma_H(r,t)-\gamma_p(r,t)| = o(1). \tag{3.13}$$

Erreur d'échantillonnage

Pour établir la convergence en probabilité vers 0 de $n(\hat{\gamma}_H(r,t)-\gamma_H(r,t))$ quand $N\to\infty$ il suffit de montrer que, pour tout $(r,t)\in[0,T]^2$

$$n^2\mathbb{E}_p\left[(\hat{\gamma}_H(r,t)-\gamma_H(r,t))^2\right]\to 0, \qquad \text{quand } N\to\infty.$$

Après avoir remarqué que

$$\begin{aligned}n|\hat{\gamma}_H(r,t)-\gamma_H(r,t)| &\le \frac{n}{N^2}\left|\sum_U\left(\frac{\hat{d}(\pi)}{d(\pi)}-1\right)\frac{\mathbb{1}_k}{\pi_k^2}(1-\pi_k)Y_k(t)Y_k(r)\right|\\ &+\frac{n}{N^2}\left|\sum_U\left(\frac{\mathbb{1}_k}{\pi_k}-1\right)\frac{1-\pi_k}{\pi_k}Y_k(t)Y_k(r)\right|\\ &+\frac{n}{N^2}\frac{1}{d(\pi)}\left|\sum_U\sum_U\left(\frac{\mathbb{1}_{kl}}{\pi_k\pi_l}-1\right)(1-\pi_k)(1-\pi_l)Y_k(t)Y_l(r)\right|\\ &:= B_1(r,t)+B_2(r,t)+B_3(r,t)\end{aligned} \tag{3.14}$$

nous obtenons

$$n^2\mathbb{E}_p\left[(\hat{\gamma}_H(r,t)-\gamma_H(r,t))^2\right]\le 3\mathbb{E}_p(B_1(r,t)^2)+3\mathbb{E}_p(B_2(r,t)^2)+3\mathbb{E}_p(B_3(r,t)^2). \tag{3.15}$$

Montrons maintenant que $\mathbb{E}_p(B_1(r,t)^2)\to 0$ quand $N\to\infty$. En utilisant les hypothèses **A1**, **A2** et **A4**, le Lemme 3.3 et l'inégalité $\frac{1}{d(\pi)}\le\frac{1}{N\lambda(1-M)}$, nous avons

$$\begin{aligned}\mathbb{E}_p(B_1(r,t)^2) &\le \frac{n^2}{d(\pi)^2}\frac{1}{N^2}\mathbb{E}_p\left[(\hat{d}(\pi)-d(\pi))^2\right]\left[\frac{1}{\lambda^4 N}\sum_U|Y_k(t)|^2|Y_k(r)|^2\right]\\ &\le\left[\frac{1}{N(1-M)}\frac{n^2}{N^2}+n\max_{k\neq l}|\Delta_{kl}|\frac{n}{N}\frac{1}{N}\right]\frac{1}{\lambda^6}\left(\frac{1}{N}\sum_U|Y_k(t)|^4\right)^{1/2}\left(\frac{1}{N}\sum_U|Y_k(r)|^4\right)^{1/2}\\ &\le\frac{1}{N}C\end{aligned}$$

Ainsi $E_p(B_1(r,t)^2)\to 0$ quand $N\to\infty$.

Sous les hypothèses **A1**, **A2** et **A4**,

$$\begin{aligned}
\mathbb{E}_p(B_2(r,t)^2) &\leq \frac{n^2}{N^4}\sum_U\sum_U \frac{|\Delta_{kl}|}{\pi_k\pi_l}\frac{1-\pi_k}{\pi_k}\frac{1-\pi_l}{\pi_l}|Y_k(t)Y_k(r)Y_l(t)Y_l(r)| \\
&\leq \frac{1}{\lambda^3}\frac{1}{N}\left(\frac{n^2}{N^2}+\frac{n^2\max_{k\neq l}|\Delta_{kl}|}{N\lambda}\right)\left(\frac{1}{N}\sum_U|Y_k(t)|^4\right)^{1/2}\left(\frac{1}{N}\sum_U|Y_k(r)|^4\right)^{1/2} \\
&\leq \frac{1}{N}C
\end{aligned}$$

D'où $\mathbb{E}_p(B_2(r,t)^2)\to 0$ quand $N\to\infty$.

$$\begin{aligned}
\mathbb{E}_p(B_3(r,t)^2) &= n^2\mathbb{E}_p\Bigg[\frac{1}{N^4}\frac{1}{d(\pi)^2}\sum_{k,l\in U}\sum_{k',l'\in U}\left(\frac{\mathbb{1}_{kl}}{\pi_k\pi_l}-1\right)\left(\frac{\mathbb{1}_{k'l'}}{\pi_{k'}\pi_{l'}}-1\right) \\
&\quad .(1-\pi_k)(1-\pi_l)(1-\pi_{k'})(1-\pi_{l'})Y_k(t)Y_l(r)Y_{k'}(t)Y_{l'}(r)\Bigg] \\
&\leq \frac{n^2}{N^4}\frac{1}{d(\pi)^2}\sum_k\sum_{k'}\left|\mathbb{E}_p\left[\left(\frac{\mathbb{1}_k}{\pi_k^2}-1\right)\left(\frac{\mathbb{1}_{k'}}{\pi_{k'}^2}-1\right)\right]\right||Y_k(t)||Y_k(r)||Y_{k'}(t)||Y_{k'}(r)| \\
&\quad +\frac{2n^2}{N^4}\frac{1}{d(\pi)^2}\sum_k\sum_{k'\neq l'}\left|\mathbb{E}_p\left[\left(\frac{\mathbb{1}_k}{\pi_k^2}-1\right)\left(\frac{\mathbb{1}_{k'l'}}{\pi_{k'}\pi_{l'}}-1\right)\right]\right||Y_k(t)||Y_k(r)||Y_{k'}(t)||Y_{l'}(r)| \\
&\quad +\frac{n^2}{N^4}\frac{1}{d(\pi)^2}\sum_{k\neq l}\sum_{k'\neq l'}\left|\mathbb{E}_p\left[\left(\frac{\mathbb{1}_{kl}}{\pi_k\pi_l}-1\right)\left(\frac{\mathbb{1}_{k'l'}}{\pi_{k'}\pi_{l'}}-1\right)\right]\right||Y_k(t)||Y_l(r)||Y_{k'}(t)||Y_{l'}(r)| \\
&:= v_1+v_2+v_3
\end{aligned}$$

En utilisant les hypothèses **A1**, **A2** et **A4**, et l'inégalité $\pi_{kl}\leq\pi_k\pi_l$, nous obtenons

$$\begin{aligned}
v_1 &\leq \frac{n^2}{N^4}\frac{1}{d(\pi)^2}\sum_k\left|\mathbb{E}_p\left[\frac{\mathbb{1}_k}{\pi_k^4}-2\frac{\mathbb{1}_k}{\pi_k^2}+1\right]\right||Y_k(t)|^2|Y_k(r)|^2 \\
&\quad +\frac{n^2}{N^4}\frac{1}{d(\pi)^2}\sum_k\sum_{k'\neq k}\left|\mathbb{E}_p\left[\frac{\mathbb{1}_{kk'}}{\pi_k^2\pi_{k'}^2}-\frac{\mathbb{1}_k}{\pi_k^2}-\frac{\mathbb{1}_{k'}}{\pi_{k'}^2}+1\right]\right||Y_k(t)||Y_k(r)||Y_{k'}(t)||Y_{k'}(r)| \\
&\leq \frac{n^2}{N^2}\frac{1}{d(\pi)^2}\left[\frac{1}{N}\left(\frac{1}{\lambda^3}+\frac{2}{\lambda}+1\right)+\left(\frac{1}{\lambda^2}+\frac{2}{\lambda}+1\right)\right]\left(\frac{1}{N}\sum_k Y_k(t)^4\right)^{1/2}\left(\frac{1}{N}\sum_k Y_k(r)^4\right)^{1/2}.
\end{aligned} \tag{3.16}$$

Depuis que $d(\pi)\to\infty$ quand $N\to\infty$ nous avons $v_1\to 0$. Sous les hypothèses **A1**, **A2**, **A4** et **A5**

$$\begin{aligned}
v_3 &\leq \frac{C}{N}+\frac{n^2}{N^4}\frac{1}{d(\pi)^2}\frac{1}{\lambda^4}\max_{\substack{k\neq l\\k'\neq l'}}|\mathbb{E}_p\left[\left(\mathbb{1}_{kl}-\pi_k\pi_l\right)\left(\mathbb{1}_{k'l'}-\pi_{k'}\pi_{l'}\right)\right]|\sum_{k\neq l}\sum_{k'\neq l'}|Y_k(t)||Y_l(r)||Y_{k'}(t)||Y_{l'}(r)| \\
&\leq \frac{C}{N}+\frac{1}{d(\pi)^2}\frac{1}{\lambda^4}n^2\max_{k\neq l,k'\neq l'}|\mathbb{E}_p\left[\left(\mathbb{1}_{kl}-\pi_k\pi_l\right)\left(\mathbb{1}_{k'l'}-\pi_{k'}\pi_{l'}\right)\right]|\left(\frac{1}{N}\sum_k|Y_k(t)|^2\right)\left(\frac{1}{N}\sum_l|Y_l(r)|^2\right)
\end{aligned} \tag{3.17}$$

Par conséquent $v_3\to 0$ quand $N\to\infty$. En appliquant l'inégalité de Cauchy-Schwarz, nous obtenons $v_2\to 0$ quand $N\to\infty$. Finalement, nous avons pour tout $(r,t)\in[0,T]^2$,

$$n\mathbb{E}_p(|\hat{\gamma}_H(r,t)-\gamma_H(r,t)|)\to 0, \quad \text{quand } N\to\infty. \tag{3.18}$$

et par conséquent,

$$n\mathbb{E}_p\left\{|\,\widehat{\gamma}_{H,d}(r,t)-\gamma_p(r,t)\,|\right\}\to 0,\qquad \text{quand } N\to\infty.$$

De la convergence en probabilité ponctuelle de la fonction de variance, nous déduisons la convergence en probabilité de toutes combinaisons linéaires finies : pour tout $p\in\{1,2,...\}$, pour tout $(c_1,...,c_p)\in\mathbb{R}^p$ et pour tout $(t_1,...,t_p)\in[0,T]^p$, nous avons

$$\sum_{l=1}^{p} c_l n(\hat{\gamma}_H(t_l,t_l)-\gamma_H(t_l,t_l))\to 0 \tag{3.19}$$

en probabilité quand N tend à l'infini. D'après le Lemme de Cramer-Wold, le vecteur $n(\hat{\gamma}_H(t_1,t_1)-\gamma_H(t_1,t_1),...,\hat{\gamma}_H(t_p,t_p)-\gamma_H(t_p,t_p))$ converge en distribution vers 0 dans $\mathbb{R}^p$.

Etape 2. Suite tendue

Pour montrer que la suite $(n(\hat{\gamma}_H(t,t)-\gamma_H(t,t))_N$ est tendue dans $C[0,T]$ nous introduisons la pseudo-métrique suivante

$$d_\gamma^2(t,r)=n^2\mathbb{E}_p(|\widehat{\gamma}_H(t,t)-\gamma_H(t,t)-\widehat{\gamma}_H(r,r)+\gamma_H(r,r)|^2)$$

et nous montrons qu'il existe une constante C telle que

$$d_\gamma^2(t,r)\le C|t-r|^{2\beta}$$

pour tout $(r,t)\in[0,T]^2$. A l'aide de l'équation (3.14), nous décomposons cette distance en 3 termes.

$$\begin{aligned}
d_\gamma^2(r,t)\le 3\Big(&\mathbb{E}_p\left([B_1(t,t)-B_1(r,r)]^2\right)+\mathbb{E}_p\left([B_2(t,t)-B_2(r,r)]^2\right)\\
&+\mathbb{E}_p\left([B_3(t,t)-B_3(r,r)]^2\right)\Big)\\
:=&\,3\left(d_{B_1}^2+d_{B_2}^2+d_{B_3}^2\right).
\end{aligned}$$

A partir du Lemme 3.1 et des hypothèses **A1** et **A2**, nous obtenons

$$\begin{aligned}
d_{B_1}^2&=\frac{n^2}{N^4}\mathbb{E}_p\left[\left(\sum_k\frac{\hat{d}(\pi)-d(\pi)}{d(\pi)}\frac{\mathbb{1}_k}{\pi_k}\frac{1-\pi_k}{\pi_k}\phi_{k,k}(t,r)\right)^2\right]\\
&\le\frac{n^2}{N^2}\left(\frac{1}{\lambda}-1\right)^2\frac{1}{\lambda^4}E_p\left[\left(\frac{1}{N}\sum_U\phi_{k,k}(t,r)\right)^2\right]\\
&\le\frac{n^2}{N^2}\left(\frac{1}{\lambda}-1\right)^2\frac{1}{\lambda^4}\zeta_1|t-r|^{2\beta}\\
&\le C|t-r|^{2\beta}.
\end{aligned}\tag{3.20}$$

D'après les hypothèses **A1**, **A2** et **A6** et le lemme 3.1,

$$\begin{aligned} d_{B_2}^2 &\leq \frac{1}{\lambda^3}\left(\frac{n^2}{N^3}+\frac{n^2\max_{k\neq l}|\Delta_{kl}|}{N^2\lambda}\right)\frac{1}{N}\sum_U|\phi_{k,k}(t,r)|^2 \\ &\leq C|t-r|^{2\beta} \end{aligned} \tag{3.21}$$

et

$$\begin{aligned} d_{B_3}^2 \leq\ & \frac{n^2}{N^4}\frac{1}{d(\pi)^2}\sum_k\sum_{k'}\left|\mathbb{E}_p\left[\left(\frac{\mathbb{1}_k}{\pi_k^2}-1\right)\left(\frac{\mathbb{1}_{k'}}{\pi_{k'}^2}-1\right)\right]\right||\phi_{k,k}(t,r)||\phi_{k',k'}(t,r)| \\ &+\frac{2n^2}{N^4}\frac{1}{d(\pi)^2}\sum_k\sum_{k'\neq l'}\left|\mathbb{E}_p\left[\left(\frac{\mathbb{1}_k}{\pi_k^2}-1\right)\left(\frac{\mathbb{1}_{k'l'}}{\pi_{k'}\pi_{l'}}-1\right)\right]\right||\phi_{k,k}(t,r)||\phi_{k',l'}(t,r)| \\ &+\frac{n^2}{N^4}\frac{1}{d(\pi)^2}\sum_{k\neq l}\sum_{k'\neq l'}\left|\mathbb{E}_p\left[\left(\frac{\mathbb{1}_{k,l}}{\pi_k\pi_l}-1\right)\left(\frac{\mathbb{1}_{k'l'}}{\pi_{k'}\pi_{l'}}-1\right)\right]\right||\phi_{k,l}(t,r)||\phi_{k',l'}(t,r)| \\ &:= b_1+b_2+b_3. \end{aligned}$$

Grâce au Lemme 3.1 et aux hypothèses **A1**, **A2** et **A4**, nous obtenons

$$\begin{aligned} b_1 &\leq \frac{n^2}{N^2}\frac{1}{d(\pi)^2}\left[\frac{1}{N}\left(\frac{1}{\lambda^3}+\frac{2}{\lambda}+1\right)+\left(\frac{1}{\lambda^2}+\frac{2}{\lambda}+1\right)\right]\frac{1}{N}\sum_k|\phi_{k,k}(t,r)|^2 \\ &\leq C|t-r|^{2\beta}. \end{aligned} \tag{3.22}$$

D'après les hypothèses **A1**, **A2**, **A4** et **A5** et le Lemme 3.2,

$$\begin{aligned} b_3 &\leq \frac{C|t-r|^{2\beta}}{N}+\frac{1}{d(\pi)^2}\frac{1}{\lambda^4}n^2\max_{k\neq l,k'\neq l'}|\mathbb{E}_p\left[\left(\mathbb{1}_{kl}-\pi_k\pi_l\right)\left(\mathbb{1}_{k'l'}-\pi_{k'}\pi_{l'}\right)\right]|\left(\frac{1}{N^2}\sum_{k,l}|\phi_{k,l}(t,r)|\right)^2 \\ &\leq C|t-r|^{2\beta}. \end{aligned} \tag{3.23}$$

En appliquant l'inégalité de Cauchy-Schwarz, nous obtenons, à partir de (3.22) et (3.23), que $b_2 \leq C|t-r|^{2\beta}$ Par conséquent

$$d_{B_3}^2 \leq C|t-r|^{2\beta}. \tag{3.24}$$

Finalement, nous déduisons des inégalités (3.20), (3.21) et (3.24) que

$$d_\gamma^2(r,t) \leq C|t-r|^{2\beta}. \tag{3.25}$$

Etant donné que $\beta > 1/2$, nous déduisons du Théorème 12.3 de Billingsley (1968) que la suite $(n(\hat{\gamma}_H(t,t)-\gamma_H(t,t)))_N$ est tendue dans $C[0,T]$.

Finalement, d'après le Théorème 8.1 de Billingsley (1968), $n(\hat{\gamma}_H(t,t)-\gamma_H(t,t))$ converge en distribution vers 0 dans $C[0,T]$.

Etant donné que $n|\widehat{\gamma}_{H,d}(r,t)-\gamma_p(r,t)| = n|\widehat{\gamma}_H(r,t)-\gamma_H(r,t)|+o(1)$, nous pouvons également déduire que $n(\widehat{\gamma}_{H,d}(t,t)-\gamma_p(t,t))$ converge en distribution vers 0 dans $C[0,T]$ (cf. Théorème 4.1 de Billingsley (1968)).

La fonctionnelle $\phi(f) = \sup_t|f(t)|$ étant continue et bornée dans $C[0,T]$, nous déduisons de la Définition 1.7 de la convergence en distribution dans $C[0,T]$ le résultat annoncé.

3.4.4 Preuve de la Proposition 3.5

Nous remarquons tout d'abord que l'erreur d'interpolation, majorée par (3.9), satisfait

$$n^{3/2}|\hat{\gamma}_{H,d}(r,t) - \hat{\gamma}_H(r,t)| = O(1) \tag{3.26}$$

à condition que $\max_{i=\{1,..,D_N-1\}} |t_{i+1} - t_i|^{2\beta} = O(n^{-1})$. Nous utilisons ensuite le fait que pour le tirage réjectif (cf. Théorème 1 de Boistard *et al.* (2012)) c_{kl} définit dans (3.11) satisfait pour une certaine constante C,

$$\max_{k,l} |c_{kl}| \leq Cd(\pi)^{-1}.$$

Ainsi, la limite (3.12) est maintenant égale à

$$d(\pi)^2 \ |\gamma_H(r,t) - \gamma_p(r,t)| = O(1). \tag{3.27}$$

Examinons maintenant l'erreur d'échantillonnage. Le seul terme de l'équation (3.15) qui n'est pas d'ordre n^{-1} est le terme $\mathbb{E}_p(B_3(r,t)^2)$. Plus précisément, en tenant compte de la majoration (3.17) nous obtenons, grâce à la Proposition 3.3 que le terme b_3 satisfait $b_3 = O(d(\pi)^{-1})$. Ainsi $\mathbb{E}_p(B_3(r,t)^2) = O(d(\pi)^{-1})$,

$$n^2\mathbb{E}_p\left[(\hat{\gamma}_H(r,t) - \gamma_H(r,t))^2\right] = O(n^{-1}) + O(d(\pi)^{-1})$$

et la preuve est terminée.

Chapitre 4

Comparaison de deux méthodes de construction de bandes de confiance

Pour mesurer la qualité d'un estimateur, il est courant, dans les enquêtes par sondage, de fournir un intervalle de confiance. Toutefois, lorsque nous travaillons avec des données fonctionnelles la question de la construction de bandes de confiance n'a été que peu abordée dans la littérature. Dans ce chapitre, nous allons proposer deux méthodes de construction. La première est basée sur la simulation de processus gaussien afin d'approcher la loi du supremum. Cette méthode s'inspire de techniques basées sur l'estimation de la fonction de covariance de l'estimateur (Faraway (1997), Cuevas *et al.* (2006) ou plus récemment Degras (2011)). Une justification asymptotique de la validité de ces techniques, dans le cadre des populations finies, est donnée dans Cardot *et al.* (2012) lorsque les hypothèses du théorème central limite sont vérifiées et que l'on dispose d'un estimateur précis de la fonction de covariance. Une deuxième méthode de construction, qui repose sur les techniques de bootstrap, adaptées aux populations finies (Booth *et al.* (1994), Chauvet (2007)) est également mise en œuvre.

Dans la première section, nous allons définir la notion de bande de confiance d'une courbe moyenne. Dans la section 4.2, nous présenterons la méthode par simulation de processus gaussien et nous donnerons une justification asymptotique pour les plans à forte entropie et pour l'estimateur basé sur un modèle de régression fonctionnelle. Dans la section 4.3, nous détaillerons les algorithmes Bootstrap pour différents plans de sondage. La section 4.4 propose enfin une comparaison de ces différentes stratégies, en termes de précision des estimateurs, de largeur et de couverture des bandes de confiance et de temps de calcul, sur l'estimation des courbes de charge d'EDF.

4.1 Introduction

Considérons, comme précédemment, une population finie $U = \{1, ..., N\}$ de taille N et supposons que, pour chaque élément k de la population U, nous pouvons observer la courbe déterministe $Y_k = (Y_k(t))_{t\in[0,T]}$. Soit s un échantillon de taille fixée n, choisi

aléatoirement dans U selon un plan de sondage $p(\cdot)$. Soient π_k =Pr$(k \in s) > 0$ et π_{kl} =Pr$(k\&l \in s)$ les probabilités d'inclusion d'ordre un et respectivement deux.

Supposons que la courbe moyenne $\mu(t)$ (cf. équation (1.37)) est estimée par l'estimateur de Horvitz-Thompson fonctionnel (cf. équation (1.38)) et que sa fonction de covariance $\gamma_p(r,t)$ (cf. équation (1.39)) est estimée par $\hat{\gamma}_p(r,t)$ (cf. équation (1.40)), pour $r,t \in [0,T]$x$[0,T]$.

Nous considérons, comme Faraway (1997) et Degras (2011), des bandes de confiance pour la courbe moyenne μ qui sont de la forme suivante

$$\mathbb{P}\left(\mu(t) \in \left[\widehat{\mu}(t) \pm c_\alpha \frac{\widehat{\sigma}(t)}{\sqrt{n}}\right],\ \forall t \in [0,T]\right) = 1-\alpha, \tag{4.1}$$

où la valeur du cœfficient c_α est inconnue, et dépend du niveau de confiance $1-\alpha$ souhaité, et $\widehat{\sigma}(t) = \sqrt{n\hat{\gamma}_p(t,t)}$.

Le calcul de c_α est basé sur le fait que, sous certaines hypothèses (Cardot *et al.* (2012)), le processus

$$\sqrt{n}\,\{\widehat{\mu}-\mu\} \to Z \text{ en distribution dans } C[0,T,] \tag{4.2}$$

où Z est un processus gaussien à valeur dans l'espace des fonctions continues $C[0,T]$.

On a alors

$$\mathbb{P}\left(\sup_{t\in T}\sqrt{n}\frac{|\hat{\mu}(t)-\mu(t))|}{\hat{\sigma}(t)} \le c_\alpha\right) \;=\; \mathbb{P}\left(\mu(t) \in \left[\widehat{\mu}(t) \pm c_\alpha \frac{\widehat{\sigma}(t)}{\sqrt{n}}\right],\ \forall t \in [0,T]\right) \tag{4.3}$$

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

Nous proposons deux approches pour déterminer la valeur de c_α. La première repose sur une estimation directe de l'écart-type et la simulation des processus gaussiens $\hat{\tilde{Z}}(t) = \sqrt{n}\frac{\hat{\mu}(t)-\mu(t)}{\hat{\sigma}(t)}$. La seconde, qui ne nécessite pas de disposer d'estimateur de la variance, repose sur des techniques de ré-échantillonnage où à la fois l'écart-type $\tilde{\sigma}(t) = \sqrt{\gamma_p(t,t)}$ et la valeur de c_α sont obtenus à partir des réplications bootstrap.

4.2 Construction de bandes de confiance par simulations de processus gaussiens

Construction de bandes de confiance par simulation de processus gaussien

Etape 1. Tirer l'échantillon s de taille n à l'aide du plan de sondage p et calculer l'estimateur $\hat{\mu}$ ainsi que l'estimateur $\widehat{\gamma}_p(r,t)$ de la fonction de covariance $\gamma_p(r,t)$, $r,t \in [0,T]$.

Etape 2. Simuler M courbes $\hat{\tilde{Z}}_m$, $m = 1\ldots,M$, de même loi que $\hat{\tilde{Z}}$ où $\hat{\tilde{Z}}$ est un processus gaussien d'espérance 0 et de fonction de covariance ρ, où $\rho(r,t) = \hat{\gamma}_p(r,t)/(\hat{\gamma}_p(t)\hat{\gamma}_p(r))^{1/2}$, $r,t \in [0,T]$.

Etape 3. Déterminer c_α, le quantile d'ordre $1-\alpha$ des variables,

$$\left(\sup_{t\in[0,T]} |\hat{\tilde{Z}}_m(t)|\right)_{m=1,\ldots,M}.$$

Cet algorithme, très rapide et facile à mettre en œuvre, a déjà été proposé, dans le cadre d'observations i.i.d. par Faraway (1997), Cuevas *et al.* (2006) et Degras (2011) pour construire des bandes de confiance. Cardot *et al.* (2012) ont donné une justification asymptotique de cette approche pour l'échantillonnage dans des populations finies de courbes bruitées.

Lorsqu'on estime la courbe moyenne à l'aide de l'estimateur assisté par un modèle de régression fonctionnelle nous pouvons également utiliser l'algorithme par simulation de processus gaussiens pour construire une bande de confiance. Il suffit de remplacer $\hat{\mu}$ par $\hat{\mu}_{MA,d}$ (cf. équation (2.11)) et $\hat{\gamma}_p$ par $\hat{\gamma}_{MA,d}$ (cf. équation (2.16)).

Proposition 4.1. *Supposons que les hypothèses* **B1-B8** *sont vérifiées et que le schéma de discrétisation satisfait* $\max_{i=\{1,..,D_N-1\}} |t_{i+1}-t_i|^{2\beta} = o(n^{-1})$.

Soit Z un processus gaussien de moyenne 0 et de fonction de covariance γ_Z (comme dans la Proposition 2.4). Soit $(\widehat{Z}_N)$ une suite de processus telle que pour chaque N, conditionnellement à l'estimateur $\widehat{\gamma}_{MA,d}$ défini dans (2.16), $\widehat{Z}_N$ est un processus gaussien de moyenne 0 et de covariance $n\widehat{\gamma}_{MA,d}$. Supposons que $\gamma_Z(t,t)$ est une fonction continue et que $\inf_t \gamma_Z(t,t) > 0$. *Alors, quand $N \to \infty$, la convergence suivante est vérifiée uniformément en c,*

$$\mathbb{P}\left(|\widehat{Z}_N(t)| \le c\widehat{\sigma}(t),\ \forall t\in[0,T] \,\middle|\, \widehat{\gamma}_{MA,d}\right) \to \mathbb{P}\left(|Z(t)| \le c\,\sigma(t),\ \forall t\in[0,T]\right),$$

où $\widehat{\sigma}(t) = \sqrt{n\widehat{\gamma}_{MA,d}(t,t)}$ *et* $\sigma(t) = \sqrt{\gamma_Z(t,t)}$.

Comme dans Cardot *et al.* (2012), nous déduisons de la proposition précédente que la valeur choisie $\widehat{c}_\alpha = c_\alpha(\widehat{\gamma}_{\mathrm{MA},d})$ fournit asymptotiquement la couverture souhaitée car elle satisfait

$$\lim_{N\to\infty} \mathbb{P}\left(\mu(t) \in \left[\widehat{\mu}_{\mathrm{MA},d}(t) \pm \widehat{c}_\alpha \frac{\widehat{\sigma}(t)}{\sqrt{n}}\right],\ \forall t\in[0,T]\right) = 1-\alpha.$$

Démonstration.

La preuve consiste à montrer la convergence faible de la suite de distribution $(\widehat{Z}_N)$ vers la loi de Z dans $C([0,T])$.

Pour tout vecteur de p instants $0 \le t_1 < \ldots < t_p \le T$, la convergence en dimension finie du vecteur de distribution gaussienne vers la distribution de $(Z(t_1),\ldots,Z(t_p))$ est une conséquence immédiate de la convergence uniforme de la fonction de variance vue dans la Proposition 2.3. Nous pouvons conclure avec le Lemme de Slutsky que pour tout $(c_1,\ldots,c_p) \in \mathbb{R}^p$,

$$\sum_{j=1}^{p}\sum_{\ell=1}^{p} c_j c_\ell \widehat{\gamma}_{\mathrm{MA},d}(t_j,t_\ell) \;\to\; \sum_{j=1}^{p}\sum_{\ell=1}^{p} c_j c_\ell \gamma_{\mathrm{MA}}(t_j,t_\ell) \quad \text{en probabilité.} \tag{4.4}$$

Maintenant, nous devons vérifier que la suite $(\widehat{Z}_N)_N$ est tendue dans $C([0,T])$. Etant donné $\widehat{\gamma}_{\mathrm{MA},d}$, nous avons pour $(r,t) \in [0,T]^2$,

$$\begin{aligned}
\mathbb{E}_p\left[\left(\widehat{Z}_N(t)-\widehat{Z}_N(r)\right)^2 \mid \widehat{\gamma}_{\mathrm{MA},d}\right] &= n\left(\widehat{\gamma}_{\mathrm{MA},d}(t,t)-2\widehat{\gamma}_{\mathrm{MA},d}(r,t)+\widehat{\gamma}_{\mathrm{MA},d}(r,r)\right)\\
&\le \frac{n}{N^2}\sum_{k,l\in U}\frac{\Delta_{kl}}{\pi_k\pi_l}\frac{\mathbb{1}_{kl}}{\pi_{kl}}\left(Y_k(t)-\widehat{Y}_{k,d}(t)-Y_k(r)+\widehat{Y}_{k,d}(r)\right)^2\\
&\le \frac{C}{N}\sum_{k\in U}\left[(Y_k(t)-Y_k(r))^2+\left(\widehat{Y}_{k,d}(t)-\widehat{Y}_{k,d}(r)\right)^2\right]\\
&\le CC_3|t-r|^{2\beta}+\frac{C}{N}\sum_{k\in U}\left(\widehat{Y}_{k,d}(t)-\widehat{Y}_{k,d}(r)\right)^2 \qquad (4.5)
\end{aligned}$$

sous les hypothèses **B2** et **B3**. Le second terme à droite de l'inégalité (4.5) est traité avec des arguments similaires à ceux de la preuve du Lemme 2.3, ainsi

$$\frac{1}{N}\sum_{k\in U}\left(\widehat{Y}_{k,d}(t)-\widehat{Y}_{k,d}(r)\right)^2 \le C|t-r|^{2\beta}.$$

Les trajectoires du processus gaussien sont continues puisque $\beta > 0$ (cf. Théorème 1.4.1 dans Adler et Taylor (2007)) et $(\widehat{Z}_N)$ converge faiblement vers Z dans $C([0,T])$ équipé de la norme *sup*. En utilisant à nouveau la Proposition 2.3, nous avons, uniformément en t, $\widehat{\sigma}_Z(t) = \sigma_Z(t) + o_p(1)$, où $\widehat{\sigma}_Z^2(t) = n\widehat{\gamma}_{\mathrm{MA},d}(t,t)$. Puisque par hypothèse $\sigma_Z^2(t) = \gamma_Z(t,t)$ est une fonction continue et $\inf_t \gamma_Z(t,t) > 0$, nous obtenons avec le Lemme de Slutsky que $(\widehat{Z}_N/\widehat{\sigma}_Z)$ converge faiblement vers Z/σ_Z dans $C([0,T])$. Par la définition de la convergence faible dans $C([0,T])$ et *the continuous mapping theorem*, nous pouvons également en déduire que la variable aléatoire $\widehat{M}_N = \sup_{t\in[0,T]} |\widehat{Z}_N(t)|/\widehat{\sigma}_Z(t)$ converge en distribution vers $M = \sup_{t\in[0,T]} |Z(t)|/\sigma_Z(t)$, de manière que pour chaque $c \ge 0$,

$$\mathbb{P}\left(\sup_{t\in[0,T]} |\widehat{Z}_N(t)|/\widehat{\sigma}_Z(t) \le c\right) \;\to\; \mathbb{P}\left(\sup_{t\in[0,T]} |Z(t)|/\sigma_Z(t) \le c\right).$$

Notons enfin, que sous les hypothèses précédentes, la variable aléatoire

$$M = \sup_{t\in[0,T]} (|Z(t)|/\sigma_Z(t))$$

a une fonction de densité absolument continue et bornée (cf. Pitt et Tran (1979)) de sorte que la convergence est uniforme en c (cf. Lemme 2.11, Van der Vaart (1998)).

□

Lorsque nous utilisons un plan de sondage à forte entropie et l'estimateur de Horvitz-Thompson, il est également possible de justifier de manière rigoureuse l'utilisation de cet algorithme pour construire une bande de confiance de notre courbe estimée $\hat{\mu}$ définie par (1.38). Nous supposons, cette fois-ci, que la covariance est estimée par $\hat{\gamma}_{H,d}(r,t)$ (cf. équation (3.5)).

Proposition 4.2. *Supposons que les hypothèses* **A1-A6** *sont vérifiées et que le schéma de discrétisation satisfait* $\max_{i=\{1,..,D_N-1\}} |t_{i+1}-t_i|^{2\beta} = o(n^{-1})$.

Soit Z un processus gaussien de moyenne 0 et de fonction de covariance $\gamma_Z = \lim_{n\to\infty} n\gamma_{p_N}(r,t)$. Soit $(\widehat{Z}_N)$ une suite de processus tel que pour tout N, conditionnellement à $\widehat{\gamma}_{H,d}$ défini dans (3.5), $\widehat{Z}_N$ est un processus gaussien de moyenne 0 et de fonction de covariance $n\widehat{\gamma}_{H,d}$. Alors pour tout $c > 0$, quand $N \to \infty$, la convergence suivante est vérifiée en probabilité

$$\mathbb{P}\left(|\widehat{Z}_N(t)| \le c\widehat{\sigma}(t),\ \forall t \in [0,T] \,\middle|\, \widehat{\gamma}_{H,d}\right) \to \mathbb{P}\left(|Z(t)| \le c\sigma(t),\ \forall t \in [0,T]\right),$$

où $\widehat{\sigma}(t) = \sqrt{n\widehat{\gamma}_{H,d}(t,t)}$ et $\sigma(t) = \sqrt{\gamma_Z(t,t)}$.

Démonstration. La preuve est similaire à celle de la proposition 4.1 et sera donc omise. □

4.3 Construction de bandes de confiance : algorithme général du bootstrap

Dans ce travail, nous avons décidé d'étendre au cadre fonctionnel le bootstrap sans remise proposé par Gross (1980). Cette méthode utilise l'échantillon s pour construire une population fictive U^* dans laquelle nous sélectionnons les échantillons bootstrapés s^*. Ces échantillons nous permettrons d'estimer la variance de l'estimateur de la moyenne en chaque instant t puis le coefficient c_α. Dans un premier temps, nous allons donner un algorithme général du bootstrap lorsque $1/\pi_k$ est entier, pour tout $k \in U$. Lorsque cette condition n'est pas vérifiée, nous adaptons ensuite cette algorithme au sondage aléatoire simple sans remise à l'aide de la variante de Booth *et al.* (1994) et au plan stratifié et au plan πps à l'aide des extensions proposées par Chauvet (2007).

Dans les algorithmes qui vont suivre, nous supposons que les trajectoires sont observées en tout point de $t \in [0,T]$.

Construction de bande de confiance par bootstrap

Etape 1. Tirer un échantillon s de taille n à l'aide du plan de sondage p et calculer l'estimateur $\hat{\mu}$.

Etape 2. Dupliquer chaque individu $k \in s$, $1/\pi_k$ fois pour construire une population fictive U^*.

Etape 3. Tirer M échantillons s_m^*, $m = 1, \ldots, M$, de taille n dans la population fictive U^* à l'aide du plan de sondage p et calculer $\hat{\mu}_m^*(t) = \frac{1}{N} \sum_{k \in s_m^*} \frac{Y_k(t)}{\pi_k}$, $t \in [0, T]$.

Etape 4. Estimer la fonction $\tilde{\sigma}(t)$ par l'écart type empirique corrigé des $\hat{\mu}_m^*(t)$, $m = 1, ..., M$,

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

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

Etape 5. Choisir c_α comme le quantile d'ordre $1 - \alpha$ des variables

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

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

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

Sondage aléatoire simple sans remise (SRSWOR)

Algorithme du bootstrap adapté au plan SRSWOR

Etape 1. Tirer un échantillon s de taille n par sondage aléatoire simple sans remise et calculer l'estimateur $\hat{\mu}_{\text{srswor}}$ défini par (1.41).

Etape 2. Dupliquer chaque individu $k \in s$, $[1/\pi_k]$ fois, où $[.]$ désigne la partie entière. On complète la population ainsi obtenue en sélectionnant un échantillon aléatoire simple sans remise dans s de taille $N - n[N/n]$. Nous obtenons ainsi une population fictive U^* de taille N.

Etape 3. Tirer l'échantillon s^* de taille n dans la population fictive U^* par sondage aléatoire simple sans remise et calculer $\hat{\mu}^*(t) = \frac{1}{n} \sum_{k \in s^*} Y_k(t)$, $t \in [0, T]$.

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

Etape 5. Répéter les étapes 4 et 5 de l'algorithme général du bootstrap.

Sondage stratifié avec SRSWOR dans chaque strate (STRAT)

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

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

$$\widehat{\mu}_{\text{strat}}(t) = \frac{1}{N} \sum_{h=1}^{H} N_h \hat{\mu}_h(t) = \sum_{h=1}^{H} \frac{N_h}{N} \left(\frac{1}{n_h} \sum_{k \in s_h} Y_k(t) \right), \ t \in [0, T]. \tag{4.6}$$

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

$$\widehat{\gamma}_{\text{strat}}(r, t) = \frac{1}{N^2} \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} \ S_{Y(r)Y(t), s_h} \ \ r, t \in [0, T], \tag{4.7}$$

où $S_{Y(r)Y(t), s_h} = \frac{1}{n_h - 1} \sum_{k \in s_h} (Y_k(r) - \hat{\mu}_h(r))(Y_k(t) - \hat{\mu}_h(t))$ est l'estimateur de la fonction de covariance $S_{Y(r)Y(t), U_h}$ dans la strate h. Pour $r = t \in [0, T]$, on obtient l'estimateur de la fonction de variance comme suit

$$\widehat{\gamma}_{\text{strat}}(r) = \frac{1}{N^2} \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} \ S^2_{Y(r), s_h}, \tag{4.8}$$

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

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

$$n_h \ = \ n \frac{N_h \sqrt{\int_0^T S^2_{Y(r), U_h} dr}}{\sum_{h=1}^{H} N_h \sqrt{\int_0^T S^2_{Y(r), U_h} dr}}, \quad h = 1, \ldots, H, \tag{4.9}$$

permettent de rendre minimale la variance intégrée de l'estimateur stratifié, $\int_0^T \widehat{\gamma}_{\text{strat}}(t) dt$. Cette allocation est similaire à l'allocation obtenue dans le cadre multivarié par Cochran (1977).

La valeur des écarts-types $S^2_{Y(r), U_h}$ de la variable d'intérêt est souvent indisponible. Si nous disposons d'une variable X connue sur toute la population et très corrélée à notre variable d'intérêt nous pouvons utiliser l'allocation x-optimale pour déterminer

la taille n_h des échantillons s_h. Ainsi, lorsque la variable X est fonctionnelle nous obtenons

$$n_h = n\frac{N_h\sqrt{\int_0^T S^2_{X(r),U_h}dr}}{\sum_{h=1}^H N_h\sqrt{\int_0^T S^2_{X(r),U_h}dr}}, \quad h = 1,\ldots,H, \tag{4.10}$$

et lorsque la variable X est univariée,

$$n_h = n\frac{N_h\sqrt{S^2_{X,U_h}}}{\sum_{h=1}^H N_h\sqrt{S^2_{X,U_h}}}, \quad h = 1,\ldots,H. \tag{4.11}$$

L'utilisation du bootstrap sans remise ne pose à priori pas de difficulté pour un échantillon stratifié, il suffit d'appliquer, indépendamment sur chaque strate, l'algorithme utilisé pour le plan SRSWOR.

Algorithme du bootstrap adapté au plan STRAT

Etape 1. Tirer un échantillon s à l'aide d'un sondage aléatoire simple sans remise de taille n_h dans chaque strate U_h, $h = 1,...,H$, et calculer l'estimateur $\hat{\mu}_{\text{strat}}$.

Etape 2. Pour $h = 1,...,H$, dupliquer chaque individu $k \in s_h$, $[N_h/n_h]$ fois, où $[.]$ désigne la partie entière. On complète les unités ainsi obtenues en sélectionnant un échantillon de taille $N_h - n_h[N_h/n_h]$ dans s_h par sondage aléatoire simple sans remise. Nous obtenons ainsi la population fictive de la strate U_h^* de taille N_h. La population fictive est $U^* = \bigcup_{h=1}^H U_h^*$ de taille N.

Etape 3. Tirer l'échantillon s^* de taille n dans la population fictive U^* selon un plan STRAT et calculer $\hat{\mu}^*(t) = \frac{1}{N}\sum_{h=1}^H \sum_{k\in s_h^*} \frac{N_h}{n_h} Y_k(t)$, $t \in [0,T]$.

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

Etape 5. Répéter les étapes 4 et 5 de l'algorithme général du bootstrap.

Plan à probabilités inégales sans remise (πps)

Lorsqu'il existe une variable auxiliaire X à valeur positive très corrélée à la variable d'intérêt nous pouvons utiliser un plan de sondage proportionnelle à la taille pour estimer la courbe moyenne μ. Dans ce cas, les probabilités d'inclusion d'ordre un sont définies par

$$\pi_k = n\frac{x_k}{\sum_U x_k}, \quad \forall k \in U. \tag{4.12}$$

Les inverses de ces probabilités étant rarement des nombres entiers, nous avons proposé d'adapter l'algorithme 3.1 proposé par Chauvet (2007).

Algorithme du bootstrap adapté au plan πps

Etape 1. Tirer un échantillon s de taille n selon le plan de sondage p proportionnel à la taille avec les probabilités d'inclusion π_k proportionnelles à x_k données par (4.12) et calculer l'estimateur $\hat{\mu}_{\pi\text{ps}}$ défini par (1.38).

Etape 2. Dupliquer chaque individu $k \in s$ $[1/\pi_k]$ fois, où $[.]$ désigne la partie entière. On complète les unités ainsi obtenues en sélectionnant un échantillon à l'aide du plan de sondage p et en prenant comme probabilité d'inclusion $\alpha_k = 1/\pi_k - [1/\pi_k]$, dans s. On obtient ainsi la population fictive U^* de taille N^*.

Etape 3. Tirer l'échantillon s^* de taille n dans U^* à l'aide du plan de sondage proportionnel à la taille avec les probabilités d'inclusion

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

et calculer $\hat{\mu}^*(t) = \frac{1}{N}\sum_{k\in s^*}\frac{Y_k(t)}{\pi_k^*}$, $t \in [0,T]$.

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

Etape 5. Répéter les étapes 4 et 5 de l'algorithme général du bootstrap.

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

Remarque 4.1. *Pour tirer des échantillons selon un plan πps, nous pouvons, par exemple, utiliser l'algorithme du cube équilibré sur la variable $\boldsymbol{\pi} = (\pi_1,...,\pi_N)$. Pour obtenir un plan proche de l'entropie maximale, il est nécessaire de trier aléatoirement la population U (resp. U^*) avant d'effectuer le tirage de l'échantillon s (resp. s_m^*) (Chauvet (2007)). D'autres algorithmes de tirage ont été proposés par Tillé (2006).*

Sondage aléatoire simple sans remise avec l'estimateur assisté par un modèle de régression fonctionnelle

On considère le modèle de superpopulation ξ présenté dans la section 2.1, c'est-à-dire

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

où $\boldsymbol{\beta}(t) = (\beta_1(t),\ldots,\beta_p(t))'$ est le vecteur des coefficients de régression fonctionnels, ϵ_{kt} sont indépendants, $\mathbb{E}_\xi(\epsilon_k) = 0$, de fonction de covariance $\text{Cov}_\xi(\epsilon_{kt},\epsilon_{kr}) = \Gamma(t,r)$ si $k = l$ et 0 sinon, pour $(t,r) \in [0,T]\times[0,T]$.

Il est également possible d'adapter l'algorithme général pour estimer la fonction de variance de l'estimateur $\hat{\mu}_{MA}$ défini par (2.5).

Algorithme du bootstrap avec l'estimateur assisté par un modèle en inférant par rapport au plan

---

Etape 1. Tirer un échantillon s de taille n par sondage aléatoire simple sans remise et calculer l'estimateur $\hat{\mu}_{MA}$ défini par (2.5).

Etape 2. Dupliquer chaque individu $k \in s$, $[1/\pi_k]$ fois, où $[.]$ désigne la partie entière. On complète la population ainsi obtenue en sélectionnant un échantillon aléatoire simple sans remise dans s de taille $N - n[N/n]$. Nous obtenons ainsi une population fictive U^* de taille N.

Etape 3. Tirer l'échantillon s^* de taille n dans la population fictive U^* par sondage aléatoire simple sans remise et calculer

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

où $\hat{\boldsymbol{\beta}}^*(t) = (\sum_{s^*} \mathbf{x}_k^* \mathbf{x}_k^{*\prime})^{-1} \sum_{s^*} \mathbf{x}_k^* Y_k^*(t)$ et $\mathbf{x}_k^*$ et $Y_k^*(t)$ sont les valeurs de $\mathbf{x}_k$ et $Y_k(t)$ dans la pseudo-population U^*.

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

Etape 5. Répéter les étapes 4 et 5 de l'algorithme général du bootstrap.

Comme le remarquent Canty et Davison (1999) le fait d'utiliser le total de la variable $\mathbf{x}_k$ sur la population U au lieu de celui sur la pseudo-population U^* conduit à de meilleurs résultats en particulier quand cette variable présente des valeurs extrêmes.

Dans l'algorithme précédent, nous avons estimé la variance en inférant uniquement par rapport au plan de sondage. Nous allons maintenant proposer de l'estimer en inférant à la fois sur le plan de sondage et sur le modèle.

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

Algorithme du bootstrap avec l'estimateur assisté par un modèle en inférant par rapport au plan et au modèle

---

Etape 1. Tirer un échantillon s de taille n selon un plan SRSWOR et calculer l'estimateur du cœfficient de régression $\widehat{\boldsymbol{\beta}}(t) = \left(\sum_s \mathbf{x}_k\mathbf{x}_k'\right)^{-1}\sum_s \mathbf{x}_k Y_k(t)$, ainsi que les résidus estimés $\hat{\epsilon}_k(t) = Y_k(t) - \mathbf{x}_k'\widehat{\boldsymbol{\beta}}(t)$, $t \in [0,T]$ et $k \in s$. La moyenne $\mu(t)$ est estimée par $\hat{\mu}_{\mathrm{MA}}(t)$ à partir de (2.5).

Etape 2. Simuler n variables aléatoires indépendantes et identiquement distribuées $Z_1, ..., Z_n$ de moyenne 0 et de variance 1 et calculer, pour chaque individu $k \in s$, les valeurs bootstrappées de Y_k,

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

Etape 3. Poser $n' = \min\{([n^2/N]+1), n\}$ et $i = [n/n']$. L'échantillon bootstrap s^* est obtenu de la manière suivante

a. Tirer par sondage aléatoire simple sans remise dans s un échantillon s_1^* de taille n'.

b. Répéter i fois l'étape (a) de façon indépendante afin de constituer l'échantillon $s^* = s_1^* \cup ... \cup s_i^*$ de taille $n^* = n'i$.

Etape 4. Calculer $\hat{\boldsymbol{\beta}}^*(t) = \left(\sum_{s^*} \mathbf{x}_k\mathbf{x}_k'\right)^{-1}\sum_{s^*} \mathbf{x}_k Y_k^*(t)$ et construire l'estimateur assisté par le modèle dans l'échantillon s^*,

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

où $\hat{Y}_k^*(t) = \mathbf{x}_k'\hat{\boldsymbol{\beta}}^*(t)$ est la valeur bootstrappée de la prédiction $\hat{Y}_k(t)$.

Etape 5. Répéter M fois les étapes 2, 3 et 4 afin d'obtenir $\hat{\mu}_m^*(t)$, $m = 1, ..., M$.

Etape 6. Répéter les étapes 4 et 5 de l'algorithme général du bootstrap.

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

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

4.4 Etude de la courbe de consommation moyenne d'électricité

Nous disposons d'une population U composée de $N = 15069$ courbes de consommation électrique mesurées toutes les demi-heures pendant deux semaines consécutives. Nous avons $D = 336$ points de mesure pour chaque semaine et nous souhaitons estimer la courbe moyenne de consommation de la deuxième semaine. On note $\mathbf{Y}'_k = (Y_k(t_1), \ldots, Y_k(t_D))$, la consommation d'électricité de l'individu $k \in U$ mesurée la deuxième semaine et $\mathbf{X}'_k = (X_k(t_1), \ldots, X_k(t_D))$ sa consommation au cours de la première semaine. La consommation moyenne de chaque individu k durant la première semaine, $x_k = \sum_{d=1}^{D} X_k(t_d)/D$, qui est une information simple et peu coûteuse à transmettre, sera utilisée comme information auxiliaire. Cette variable est également fortement liée à la courbe de consommation courante (cf. Figure 4.1).

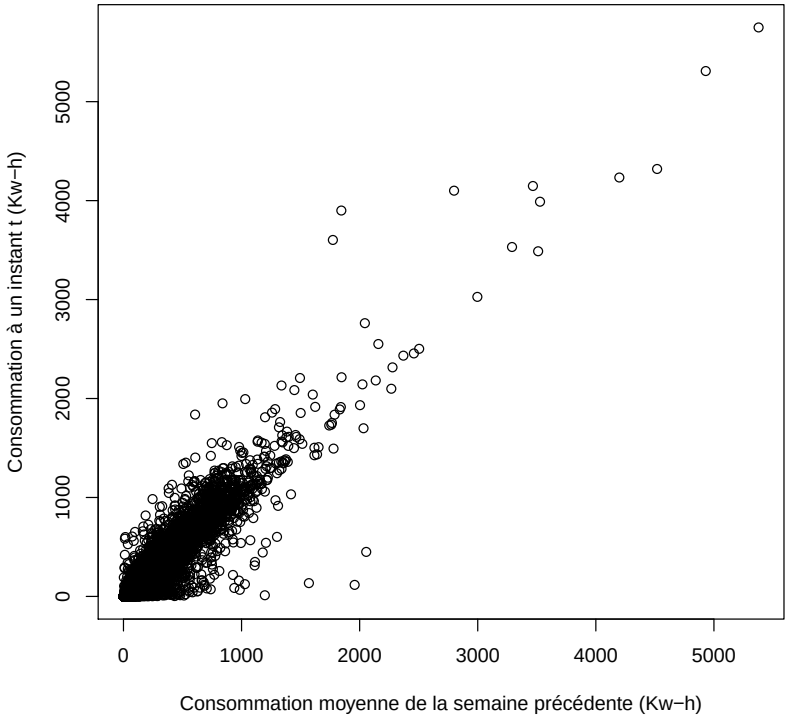


FIGURE 4.1 – Représentation de la consommation à un instant t en fonction de la consommation moyenne de la semaine précédente.

4.4.1 Description des stratégies utilisées

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

1. *Sondage SRSWOR et estimateur de Horvitz-Thompson.* La mise en œuvre de ce plan est simple, l'estimateur de Horvitz-Thompson de la courbe moyenne est

donné par (1.41) et l'estimateur de sa variance par (1.43).

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

   Les tailles des strates N_h obtenues en utilisant la première (resp. la deuxième) stratification ainsi que les tailles n_h optimales, selon (4.10) (resp. (4.11)), des échantillons à sélectionner dans chaque strate sont données dans le tableau 4.1 (resp. 4.2). Dans les deux cas, les strates sont numérotées en ordre croissant par rapport à la consommation moyenne de chaque strate. Plus précisement, la strate 1 correspond aux faibles consommateurs et la strate 10 est composée des 10 plus gros consommateurs d'électricité. Il faut remarquer aussi que la première stratification exige plus d'information que la deuxième stratification car dans le premier cas, il faut connaître la consommation d'électricité à chaque instant de mesure t. La courbe moyenne est construite en utilisant (4.6) et sa variance est estimée par (4.8).

3. *Sondage πps et estimateur de Horvitz-Thompson.* Nous avons utilisé l'algorithme du cube proposé par Deville et Tillé (2004) et Chauvet et Tillé (2006) où les probabilités d'inclusion sont proportionnelles à $x_k, k \in U$. Afin d'avoir un plan de sondage proche de l'entropie maximale, un tri aléatoire de la population est effectué avant le tirage de l'échantillon s. La covariance de l'estimateur de la moyenne est estimée à l'aide de la formule (3.5). L'algorithme du cube est disponible sous R dans le package *sampling*, fonction *samplecube* et une macro `SAS` est disponible sur le site web de l'INSEE (Institut National de Statistique et des Etudes Economiques).

4. *Sondage SRSWOR et estimateur MA.* L'estimateur $\hat{\mu}_{MA}$ assisté par le modèle ξ est construit à l'aide de l'information auxiliaire donnée par $\mathbf{x}'_k = (1, x_k)$ où x_k est la consommation moyenne de la semaine précédente. Dans ces conditions, $\hat{\mu}_{MA}$ est la somme sur toute la population U des valeurs prédites $\hat{Y}_k$ par le modèle (cf. formule (2.6)). La covariance de l'estimateur de la moyenne est estimée à l'aide de la formule (2.16).

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

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

h	1	2	3	4	5	6	7	8	9	10
N_h	3257	4236	3139	1937	1189	731	415	125	30	10
n_h	260	293	248	204	159	133	111	56	26	10

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

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

4.4.2 Erreur d'estimation de la courbe moyenne

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

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

et

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

Les résultats sont présentés dans les Tables 4.3 et 4.4 pour $I = 10000$ simulations (réplications). Ils montrent clairement que, pour cette étude, la prise en compte de la consommation totale de la semaine précédente permet d'améliorer de manière importante la précision de l'estimation de la moyenne par rapport au sondage aléatoire simple sans remise en divisant l'erreur moyenne absolue R_1 par 5/2. Parmi les différentes stratégies, les plus performantes semblent être celles qui prennent en compte l'information auxiliaire via les probabilités d'inclusion (STRAT, π-ps et systématique proportionnel à la taille).

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

TABLE 4.3 – Erreur R_1 d'estimation de la moyenne μ, avec $I = 10000$ réplications.

Stratégie	moyenne	1er quartile	médiane	3eme quartile
SRSWOR	40.53	10.82	22.16	51.09
STRAT (1)	5.78	3.68	5.08	7.07
STRAT (2)	6.49	4.03	5.48	7.88
πps	7.06	3.99	5.52	8.16
π-ps systématique	6.73	3.85	5.20	8.07
MA	8.29	5.24	7.14	10.06

TABLE 4.4 – Erreur quadratique R_2 d'estimation de la moyenne μ, avec I = 10000 réplications.

4.4.3 Taux de couverture et largeur des bandes de confiance

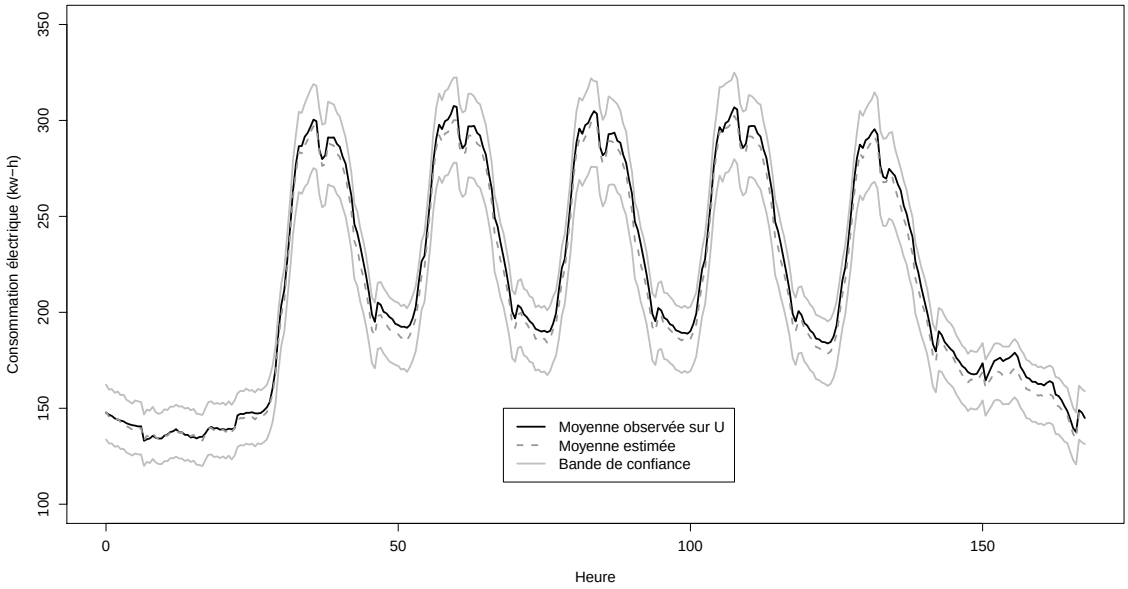
La construction des bandes de confiance de niveau $1-\alpha$ nécessite le calcul des quantiles d'ordre $1-\alpha$ du supremum de processus.

Pour ne pas privilégier une méthode de construction de bande de confiance par rapport à l'autre, nous avons appliqué les 2 algorithmes sur un même échantillon s et nous avons considéré le même nombre M de processus. Ce nombre M varie d'un estimateur à l'autre en raison des temps de calculs nécessaires pour les approches de type bootstrap (voir Section 4.4.4).

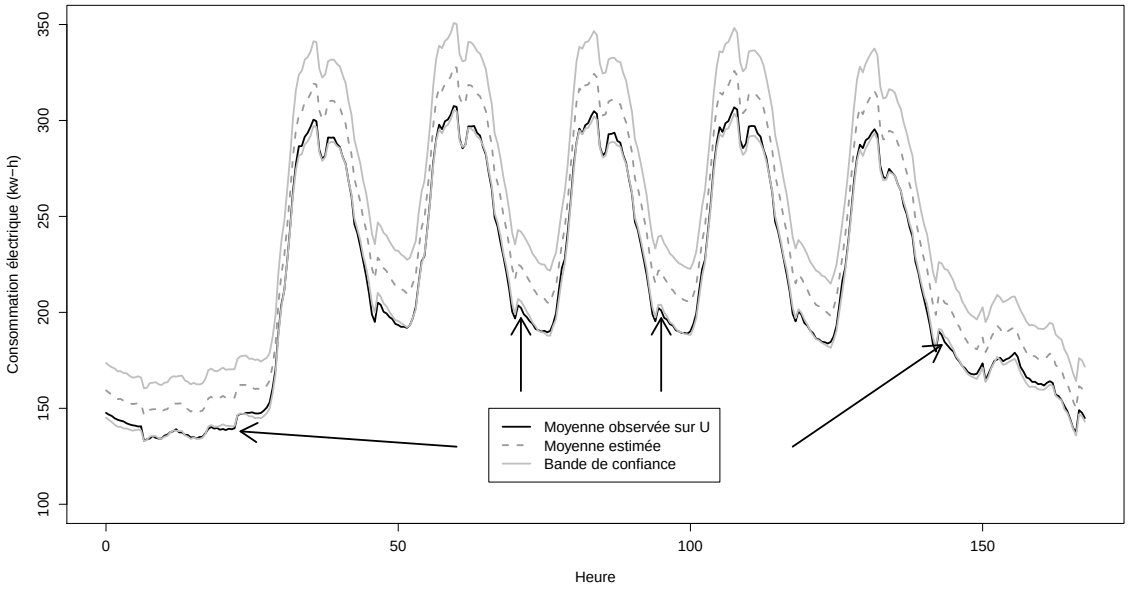
Le taux de couverture empirique est la proportion de fois, parmi les I = 2000 réplications, où la vraie courbe moyenne μ se trouve, pour tous les instants t, à l'intérieur de la bande de confiance construite à partir d'une estimation $\hat{\mu}$. Nous avons représenté sur la Figure 4.2 deux exemples de bandes de confiance (courbes grises continues) construites à partir des courbes estimées (courbes grises pointillées). Sur la Figure 4.2(a), nous constatons que la vraie courbe moyenne sur la population (courbe noir continue) est à l'intérieur de la bande de confiance à chaque instant. A l'opposé, sur la Figure 4.2(b), nous constatons que la courbe moyenne de la population est en général surestimée et qu'il existe quelques instants (indiqués par les flèches) où la courbe observée sort de la bande de confiance. Les taux de couverture empiriques sont présentés dans la Table 4.5.

Pour construire la bande de confiance de l'estimateur $\hat{\mu}_{MA}$ nous avons considéré les 2 algorithmes bootstrap décrits dans la section précédente, on notera MA (1) le bootstrap réalisé en inférant sur le plan et MA (2) celui réalisé en inférant à la fois sur le plan et sur le modèle.

Les deux méthodes de construction des bandes de confiance donnent des taux de couverture similaires et assez proches des taux nominaux souhaités (95 % et 99 %). Les résultats semblent cependant légèrement moins satisfaisants pour les plans πps et pour l'approche MA pour lesquelles la variance de l'estimateur est complexe et plus difficile à estimer précisément.



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



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

FIGURE 4.2 – Exemples de bande de confiance

Méthodes	Nombre M de processus	Bootstrap		Processus gaussien	
		α=0.05	α=0.01	α=0.05	α=0.01
SRSWOR	5000	94.95	98.85	94.80	98.70
STRAT (1)	5000	93.92	98.34	94.09	98.43
STRAT (2)	5000	94.3	98.45	94	98.55
πps	1000	94.73	98.77	93.87	98.61
MA (1)	5000	94.3	98.5	92.85	98.15
MA (2)	5000	94.4	99.05	93.15	98.70

TABLE 4.5 – Taux de couverture empirique (en %), pour I=2000 réplications.

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

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

dont les valeurs sont présentées dans la Table 4.6. Les deux méthodes fournissent des bandes de confiance dont les largeurs sont similaires. On note également que l'utilisation de la variable auxiliaire permet de diminuer sensiblement la largeur moyenne des bandes, celle-ci est divisée par 2 si on considère un des plans stratifiés plutôt qu'un plan SRSWOR.

Méthodes	Nombre M de processus	Bootstrap		Processus gaussien	
		α=0.05	α=0.01	α=0.05	α=0.01
SRSWOR	5000	35.98	43.35	35.99	43.19
STRAT (1)	5000	16.64	18.92	16.62	18.88
STRAT (2)	5000	17.58	19.99	17.55	19.94
πps	1000	17.85	20.31	17.62	19.93
MA (1)	5000	19.88	22.65	19.75	22.44
MA (2)	5000	20.03	22.93	19.72	22.40

TABLE 4.6 – Largeur moyenne des bandes de confiance, pour I = 2000 réplications.

Les Figures 4.3 et 4.4 présentent les largeurs des bandes de confiance pour un niveau α = 0.05, pour chaque instant, selon qu'elles soient ponctuelles (c_α = 1.96), estimées par simulations de processus gaussiens ou bien obtenues en considérant l'approche basée sur l'inégalité de Bonferroni appliquée en chaque point de mesure. On a alors, dans ce dernier cas, c_α = 3.793048, le quantile d'ordre $1 - 0.05/(336 \times 2)$ d'une loi $N(0,1)$. Les bandes obtenues par Bonferroni sont conservatives et considèrent en quelque sorte le pire des cas en termes d'information, celui de l'indépendance des intervalles ponctuels. On peut remarquer que l'approche par simulation permet de réduire sensiblement la largeur moyenne des bandes en comparaison avec Bonferroni lorsque le plan ne permet pas de prendre en compte toute l'information temporelle des données (Figure 4.3). A l'opposé, pour le plan stratifié (Figure 4.4) qui permet une estimation précise de la

courbe moyenne, la bande de confiance construite par simulation est proche de celle de Bonferroni, ce qui s'interprète intuitivement comme le fait que quasiment toute l'information a été capturée par le plan de sondage.

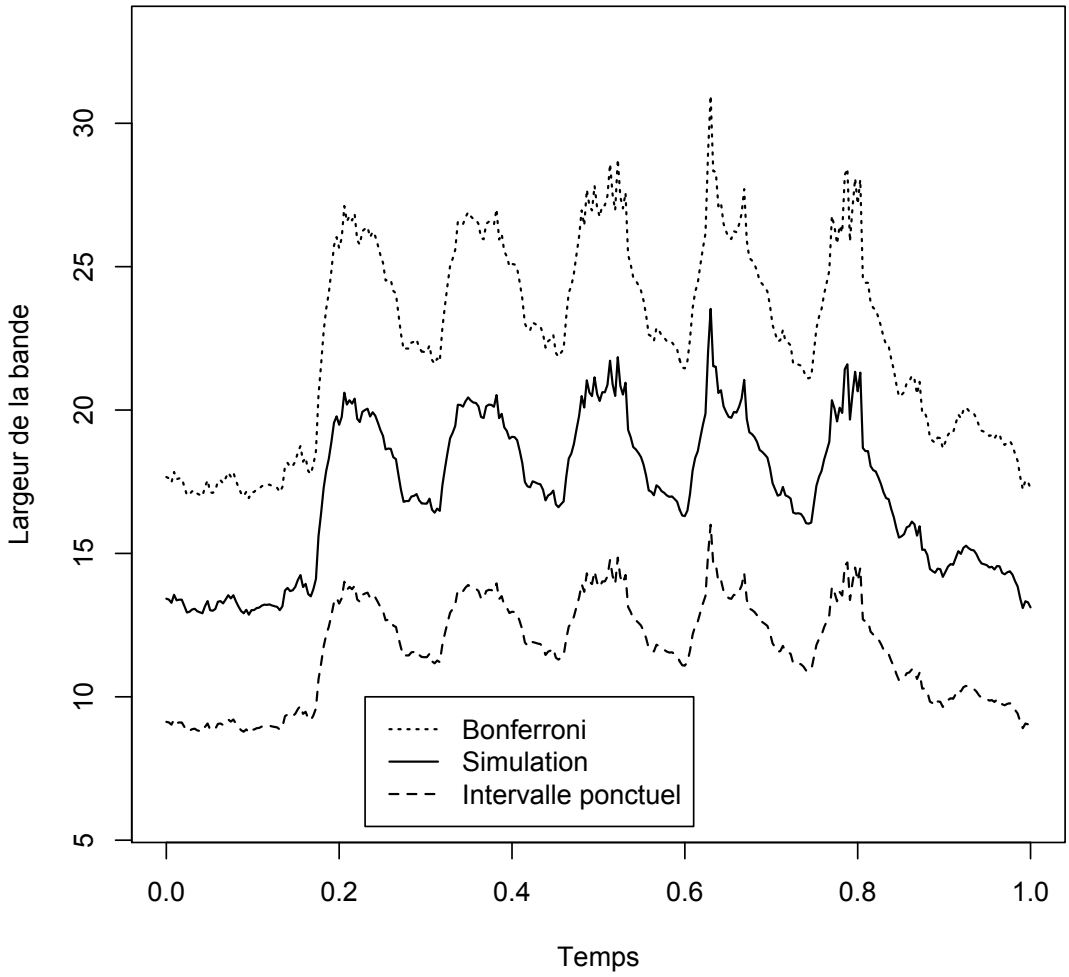


FIGURE 4.3 – Sondage aléatoire simple sans remise. Largeur des bandes de confiance ponctuelles, globales par simulations de processus et avec Bonferroni ($\alpha = 0.05$).

4.4.4 Temps de calcul

Les temps de calcul avec la méthode par bootstrap sont largement supérieurs, de l'ordre d'un facteur de 1 à 1000, à ceux de la méthode par simulations de processus gaussiens (cf. Table 4.7). Cette différence importante provient du fait que les méthodes de bootstrap nécessitent de répéter tout le processus d'estimation pour chaque échantillon bootstrapé : construction de la population fictive, tirage d'un nouvel échantillon, calcul de l'estimateur. On remarque également que les plans qui font intervenir de

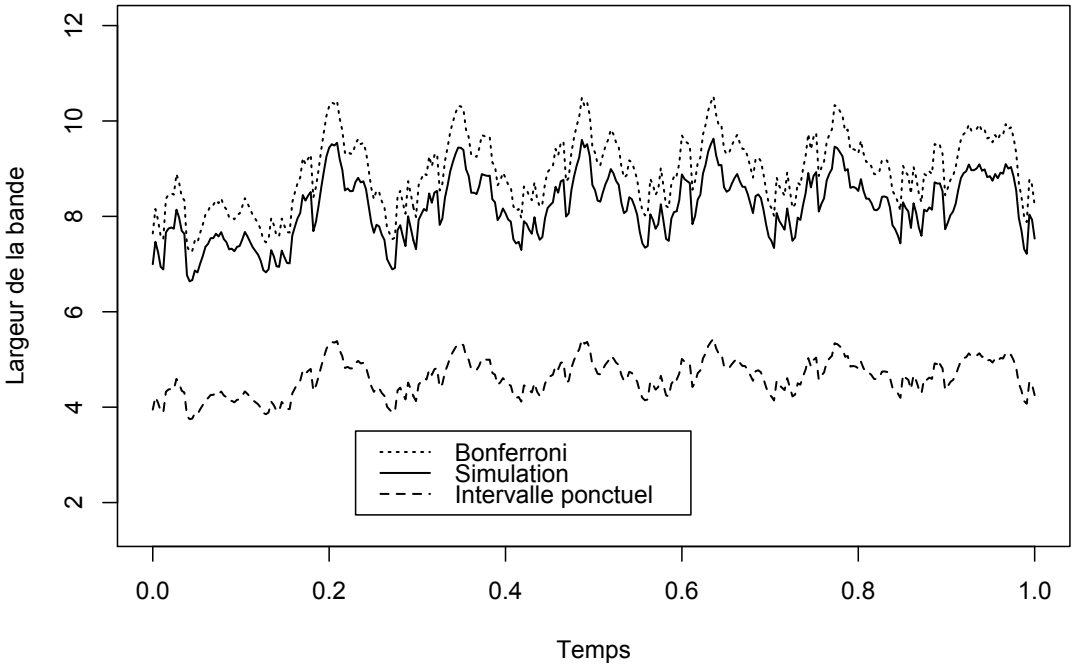


FIGURE 4.4 – Sondage stratifié (STRAT 1). Largeur des bandes de confiance ponctuelles, globales par simulations de processus et avec Bonferroni (avec $\alpha = 0.05$).

l'information auxiliaire sont moins rapides que le plan SRSWOR même si utilisés individuellement leur temps de calcul reste tout à fait raisonnable.

Stratégie	Bootstrap	Processus gaussiens
SRSWOR	1170.6	1.0
STRAT	1839.5	1.4
πps	5020.0	7.3
MA(1)	3156	1.4
MA (2)	2423.4	1.4

TABLE 4.7 – Temps d'exécution d'une simulation en secondes pour M=5000 réplications. Les stratégies SRSWOR, MA et STRAT ont été programmés avec R et πps avec SAS.

4.5 Conclusion

Nous avons, dans ce chapitre, mis en œuvre et comparé différentes stratégies permettant de prendre en compte de l'information auxiliaire pour l'estimation, et la

construction de bandes de confiance, de la moyenne de données qui sont des courbes. Cette information peut être prise en compte au moment de l'échantillonnage en considérant des plans à probabilités inégales ou bien lors de l'estimation avec un sondage aléatoire simple sans remise assisté par un modèle de régression à réponse fonctionnelle. Il apparaît clairement, sur notre exemple de courbes de charge d'électricité, que la connaissance des consommations totales une semaine avant, permet d'améliorer de manière importante la précision des estimateurs de la moyenne par rapport à un sondage de type SRSWOR.

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

Conclusion et perspectives

Dans cette thèse, nous avons cherché à améliorer la précision de l'estimation de la courbe moyenne en prenant en compte l'information auxiliaire disponible. Pour cela, nous avons développé deux méthodes d'estimation qui combinent les techniques de sondage et la statistique fonctionnelle. La première utilise un plan de sondage à probabilités inégales à forte entropie et l'estimateur de Horvitz-Thompson fonctionnel. La seconde est basée sur un modèle de régression fonctionnelle. Pour chacune de ces méthodes, nous avons démontré la convergence uniforme de l'estimateur de la moyenne et de sa fonction de covariance et nous avons également établi un théorème central limite fonctionnel. Dans le cas d'un tirage réjectif, nous avons pu établir la vitesse de convergence de l'estimateur de la covariance de l'estimateur de Horvitz-Thompson. Dans un deuxième temps, nous avons comparé deux méthodes de construction de bande de confiance sur un jeu de données de courbes de consommation électrique. La première utilise la normalité asymptotique de nos estimateurs et repose sur la simulation du processus gaussien limite en utilisant une estimation de sa variance. La seconde est basée sur des techniques de bootstrap en population finie. Celle-ci a l'avantage de ne pas utiliser la fonction de covariance estimée mais elle est beaucoup plus gourmande en temps de calcul.

Dans notre étude, nous avons constaté que notre variable auxiliaire permet d'améliorer significativement la précision de notre estimateur et que les deux méthodes de construction de bandes de confiance donnent des taux de couverture proches des taux nominaux souhaités. Cependant, pour le plan πps et l'estimateur assisté par le modèle, la variance est plus difficile à estimer et les taux de couverture sont donc légèrement inférieurs.

Ces différents travaux pourraient être améliorés et étendus dans de nombreuses directions.

Dans notre étude, la liaison entre notre variable d'intérêt et notre variable auxiliaire étant clairement linéaire (cf. Figure 4.1), nous nous sommes focalisés sur l'estimateur basé sur un modèle de régression linéaire fonctionnelle avec une variable auxiliaire réelle. Une poursuite intéressante de ces travaux serait d'étendre le modèle au cas de variables auxiliaires fonctionnelles (par exemple inclure la courbe de température). Nous pourrions également développer une méthode d'estimation basée sur des modèles non-paramétriques. Ces derniers nécessiteraient de connaitre les variables auxiliaires pour l'ensemble de la population.

Le plan πps mis en place dans nos applications permet d'obtenir une bonne estimation de la courbe moyenne de consommation et de sa covariance (cf. tableaux 3.1 et 4.3). Cependant, nous avons dans notre population certains individus dont la probabilité d'inclusion est très faible. Ces individus peuvent avoir une contribution très importante à l'estimateur et conduire à une mauvaise estimation à la fois de la courbe moyenne mais aussi de la variance de l'estimateur. Par exemple, un individu qui revient de vacances durant notre semaine d'étude aura une valeur faible de π_k et un niveau moyen de consommation très différent de celui de la semaine précédente. Nous pouvons également être dans la situation contraire, c'est-à-dire nous donnons un poids trop faible aux individus qui ont une forte probabilité d'inclusion et une diminution de leur consommation durant la semaine d'étude (exemple : départ en vacances). Il s'agit d'un problème de robustesse du plan de sondage et de l'estimateur par rapport au paramètre d'intérêt étudié. Nous pouvons envisager d'utiliser des méthodes comme celles de Beaumont et Rivest (2009), pour repérer de tels individus et corriger leur poids.

Lorsqu'on travaille sur une longue période d'étude, notre plan de sondage représente bien notre population au début de la période mais il est probable que sa représentativité varie au cours du temps (inclusion de nouveaux tarifs, nouveaux clients, etc.). Une autre piste de recherche serait donc de faire évoluer notre échantillon au cours du temps. Un premier travail réalisé par Degras (2012) montre clairement que, dans le cas du sondage stratifié et de l'estimateur de Horvitz-Thompson, la performance de notre plan de sondage est améliorée lorsque notre échantillon varie au cours du temps.

Dans toutes les études réalisées, nous avons été confrontés à la présence de courbes qui ne sont que partiellement observées dans notre population d'étude. Pour des raisons de simplicité, nous avons décidé de les exclure de notre population. Or la non-prise en compte de ces individus crée une erreur d'estimation en terme de biais et de variance de notre courbe moyenne. Différents travaux ont été réalisés pour prendre en compte la non-réponse quand le paramètre à estimer est univarié : repondération (Brick et Montequila (2009), Lundström et Särndal (1999), Särndal et Lundström (2005)), imputation (Haziza (2009)). Lorsqu'on travaille avec des données de types courbes la problématique est un peu différente. En effet, la consommation d'un individu à un instant t est fortement corrélée à sa consommation aux instants précédents et suivants. Celle-ci dépend également d'autres variables auxiliaires telles que le type d'habitation, les caractéristiques du client, etc. Pour reconstituer les trajectoires manquantes, nous pouvons appliquer, instant par instant, les méthodes classiques d'imputation. L'inconvénient de ces méthodes essentiellement univariées est qu'elles ne prennent pas en compte l'historique de consommations des individus. Notons également qu'une difficulté supplémentaire provient du fait que cet historique peut également être entaché par la non-réponse. Une deuxième possibilité est d'appliquer des méthodes d'interpolation ou de lissage des trajectoires. Cette approche permet de reconstituer, indépendamment des autres individus, les trajectoires individuelles en prenant uniquement en compte leur historique de consommation. Il faut donc trouver une méthode d'imputation qui

permette d'imputer des morceaux de courbes en prenant en compte l'ensemble des consommations observées sur notre échantillon ainsi que l'information auxiliaire. Pour cela, nous avons envisagé d'adapter au cadre sondage les travaux de Hall *et al.* (2006). Dans ces travaux, les auteurs font appel à des méthodes de lissage et à une estimation de la matrice de covariance pour estimer la courbe moyenne de consommation à partir d'observations discrétisées. Nous allons présenter en quelques lignes l'idée de la méthode que nous envisageons.

Soit s un échantillon de taille fixée n, choisi aléatoirement dans notre population d'étude U_N selon un plan de sondage $p(\cdot)$. Nous supposons que les probabilités d'inclusion du premier ordre satisfont $\pi_k = \mathbb{P}(k \in s) > 0$, pour tout $k \in U$ et que pour chaque individu de l'échantillon s, nous avons observé la consommation Y_k aux instants t_{kj}, $j = 1, ..., p_k$ (cf. Hall *et al.* (2006)).

Considérons le modèle de superpopulation suivant. Soit une variable aléatoire $X = \{X(t), t \in [0,1]\}$, d'espérance $\mathbb{E}(X(t)) = \mu(t)$ et de fonction de covariance $\mathrm{Cov}(X(s), X(t)) = \gamma(s,t)$. On considère une suite de réalisations indépendantes $X_1, \ldots, X_N, \ldots$ de la variable X. La population U_N sera alors constituée des réalisations $X_1, ..., X_N$. Nous supposons également que les observations $Y_k(t_{kj})$ satisfont pour tout $k \in s$ et $j = 1, ..., p_k$,

$$Y_k(t_{kj}) = X_k(t_{kj}) + \epsilon_{kj}, \quad k \in s, \ j = 1, \ldots, p_k.$$

où les ϵ_{kj} sont i.i.d centrés et de variances σ^2.

L'objectif est d'estimer la courbe moyenne de consommation $\mu(t)$ à partir des observations $Y_k(t_{kj})$, $k \in s$ et $j = 1, ..., p_k$ et de reconstituer les trajectoires de chaque individu de l'échantillon.

Etape 1 : estimation de la courbe moyenne $\mu(t)$

Si nous disposons des consommations Y_k pour l'ensemble des individus de notre population U_N, nous pouvons estimer la courbe moyenne par (cf. Staniswalis et Lee (1998), Hall *et al.* (2006))

$$\tilde{\mu}(t) = \frac{\sum_{k=1}^{N} \sum_{j=1}^{p_k} K(t - t_{kj}, h) Y_k(t_{kj})}{\sum_{k=1}^{N} \sum_{j=1}^{p_k} K(t - t_{kj}, h)}, \quad t \in [0,1]$$

où K est un noyau et h est la fenêtre de lissage. Les consommations étant connues uniquement pour les individus de l'échantillon, nous pouvons estimer $\mu(t)$ en remplaçant chaque somme dans $\tilde{\mu}(t)$ par son estimateur de Horvitz-Thompson

$$\widehat{\mu}(t) = \frac{\sum_{k \in s} w_k \sum_{j=1}^{p_k} K(t - t_{kj}, h) Y_k(t_{kj})}{\sum_{k \in s} w_k \sum_{j=1}^{p_k} K(t - t_{kj}, h)}, \quad t \in [0,1]$$

avec $w_k = \pi_k^{-1}$.

Etape 2 : reconstitution des trajectoires à l'aide du BLUP

Pour reconstituer les trajectoires $X_k(t)$, nous faisons appel à la formule du BLUP (Best Linear Unbiased Prediction) : c'est l'espérance conditionnelle si la loi jointe est gaussienne et c'est la meilleure approximation linéaire (au sens de l'erreur quadratique moyenne) de l'espérance conditionnelle sinon. Elle peut s'écrire

$$\mathbb{E}\left(X_k(t)|\mathbf{Y}_k\right) = \mu(t) + \sum_{j=1}^{p_k} \alpha_{kj}(t)\left(\mu(t_{kj}) - Y_k(t_{kj})\right)$$

où $\mathbf{Y}_k = (Y_k(t_{kj}), j = 1, \ldots, p_k) \in \mathbb{R}^{p_k}$ et $\boldsymbol{\alpha}_k(t) = \mathrm{Cov}(X_k(t), \mathbf{Y}_k)\left(\mathrm{Var}(\mathbf{Y}_k)\right)^{-1}$.

De nos hypothèses, nous déduisons que

$$\mathrm{Cov}(X_k(t), \mathbf{Y}_k) = \left(\gamma(t, t_{k1}), \ldots, \gamma(t, t_{kp_k})\right)$$

et

$$[\mathrm{Var}(\mathbf{Y}_k)]_{\ell_1\ell_2} = \mathrm{Cov}(Y_k(t_{k\ell_1}), Y_k(t_{k\ell_2})) = \gamma(t_{\ell_1}, t_{\ell_2}) + \sigma^2 \mathbb{1}_{\{\ell_1=\ell_2\}}.$$

Pour estimer les trajectoires, il faut donc déterminer un estimateur de la covariance $\gamma(r,t) = \mathbb{E}(X(t)X(r)) - \mu(t)\mu(r)$ et plus particulièrement de la fonction $(r,t) \mapsto \mathbb{E}(X(t)X(r))$. Si les $Y_k(t_{kj})$ étaient connus pour tous les individus de la population, on utiliserait l'estimateur suivant

$$\widehat{\mathbb{E}}(X(t)X(r)) = \frac{\displaystyle\sum_{k=1}^{N}\sum_{\ell,j,j\neq\ell}^{p_k}\left(Y_k(t_{kj}) - \tilde{\mu}(t_{kj})\right)\left(Y_k(t_{k\ell}) - \tilde{\mu}(t_{k\ell})\right)K(t-t_{kj},h)K(r-t_{k\ell},h)}{\displaystyle\sum_{k=1}^{N}\sum_{\ell,j,j\neq\ell}^{p_k}K(t-t_{kj},h)K(r-t_{k\ell},h)}.$$

En tenant compte des probabilités d'inclusion, on peut proposer l'estimateur pondéré suivant

$$\widehat{\mathbb{E}}_s(X(t)X(r)) = \frac{\displaystyle\sum_{k\in s} w_k \sum_{\ell,j,j\neq\ell}^{p_k}\left(Y_k(t_{kj}) - \hat{\mu}(t_{kj})\right)\left(Y_k(t_{k\ell}) - \hat{\mu}(t_{k\ell})\right)K(t-t_{kj},h)K(r-t_{k\ell},h)}{\displaystyle\sum_{k\in s} w_k \sum_{\ell,j,j\neq\ell}^{p_k}K(t-t_{kj},h)K(r-t_{k\ell},h)}.$$

Ainsi la covariance $\gamma(r,t)$ est estimée pour tout $(t,r) \in [0,1]\mathrm{x}[0,1]$ par

$$\hat{\gamma}(r,t) = \widehat{\mathbb{E}}_s(X(t)X(r)) - \hat{\mu}(t)\hat{\mu}(r).$$

et la trajectoire $X_k(t)$ sera alors estimée par

$$\hat{X}_k(t) = \hat{\mu}(t) + \sum_{j=1}^{p_k}\hat{\alpha}_{kj}(t)\left(\hat{\mu}(t_{kj}) - Y_k(t_{kj})\right), \quad t \in [0,1]$$

où $\boldsymbol{\alpha}_k(t) = \widehat{\mathrm{Cov}}(X_k(t), \mathbf{Y}_k)\left(\widehat{\mathrm{Var}}(\mathbf{Y}_k)\right)^{-1}$.

Si la variance du bruit σ^2 est nulle, la trajectoire estimée passera par les points observées $Y_k(t_{kj})$.

Etape 3 : estimation de la courbe moyenne à partir des trajectoires estimées

Une fois que nous avons reconstitué nos trajectoires $X_k(t)$, on peut envisager d'estimer la courbe moyenne par l'estimateur suivant

$$\hat{\mu}_{\hat{X}} = \frac{1}{N} \sum_s \frac{\hat{X}_k(t)}{\pi_k}, \quad t \in [0,1] \tag{4.13}$$

Certains points de cette méthode sont encore à développer avant de pouvoir la tester sur un jeu de données (choix du noyaux et de la fenêtre de lissage, inclusion des variables auxiliaires dans l'étape de reconstruction des trajectoires, calcul du biais et de la variance, etc.).

Annexe

Annexe A

Estimation de la courbe de consommation des particuliers à l'aide de la température extérieure

Le profil de consommation des particuliers est fortement dépendant de covariables telles que la consommation passée, les caractéristiques météorologiques ou géographiques. Nous avons vu dans le chapitre 1 qu'il existe différentes méthodes qui nous permettent de prendre en compte ces informations et ainsi d'améliorer notre estimation (diminution de la variance). Nous pouvons la faire intervenir soit au niveau du tirage de l'échantillon (sondage à probabilité inégales, équilibrés, etc.) soit au niveau de l'estimation (estimateur assisté par un modèle, calage, etc.). Le choix de la méthode va dépendre de notre information auxiliaire. Si celle-ci est disponible avant l'obtention de l'échantillon nous pourrons la faire intervenir au niveau du plan de sondage et/ou au niveau de l'estimation (consommations passées, coordonnées géographiques, etc.). Par contre, si nous ne l'obtenons qu'au moment de l'échantillonnage nous ne pourrons la faire intervenir qu'au niveau de l'estimation. Dans notre cas, la température ne sera pas connue avant le tirage. La seule manière de l'utiliser est d'estimer notre courbe de consommation moyenne avec un estimateur assisté par un modèle ou par calage.

Dans ce chapitre, nous allons apporter un début de réponse à la question suivante : est-ce que l'utilisation de variables auxiliaires telles que la température extérieure permet d'améliorer la précision de l'estimateur de la courbe moyenne de consommation des particuliers ? Pour répondre à cette question, nous avons cherché une méthode d'estimation de la courbe moyenne qui prenne en compte à la fois l'information auxiliaire disponible et le fait que nous travaillons avec des données qui peuvent être vues comme des données fonctionnelles. Plus précisément, nous utiliserons le plan de sondage SRSWOR et l'estimateur assisté par un modèle linéaire fonctionnel introduit dans le chapitre 2 ainsi que l'estimateur assisté par un modèle basé sur l'ACP fonctionnelle (Cardot *et al.* (2010b)) décrit dans la section ci-dessous. Enfin, nous utiliserons l'information auxiliaire au niveau du plan d'échantillonnage. Nous comparerons plusieurs

stratégies (plan de sondage et estimateur) qui prennent en compte la température pour estimer la courbe moyenne.

Pour des raisons de confidentialité, certains résultats de cette étude seront omis.

A.1 Estimateur assisté par un modèle basé sur l'ACP fonctionnelle

Soit $U = \{1, ..., N\}$ une population finie de taille N, dans laquelle un échantillon s de taille n est sélectionné à l'aide d'un plan de sondage $p(.)$. Nous supposons que $\pi_k = Pr(k \in s) > 0 \ \forall k \in U$ et que $\pi_{kl} = Pr(k\&l \in s) > 0 \ \forall k, l \in U$, $k \neq l$.

La courbe de consommation $Y_k = (Y_k(t))_{t\in[0,T]}$ est supposée observable pour tout individu k de la population U. L'objectif est d'estimer la courbe de consommation moyenne de la population U

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

Soient $X_1, ..., X_p$, p variables auxiliaires liées à la variable Y et observées pour chaque individu k de la population U. On note $\mathbf{x}_k = (x_{1k}, ..., x_{pk})'$ le vecteur contenant les p variables auxiliaires observées pour l'individu k. Considérons le modèle de superpopulation ξ introduit par Cardot *et al.* (2010b) défini comme suit

$$\xi : \quad Y_k(t) \quad = \quad \mu(t) + \sum_{j=1}^{q} < Y_k - \mu, v_j > v_j(t) + \eta_k(t) \tag{A.1}$$

avec $v_j(t)$ est la j^{me} fonction propre de la matrice de covariance Γ des Y_k.

Cardot *et al.* (2010b) supposent que les composantes $c_{jk} =< Y_k - \mu, v_j >$, $j = 1, ..., q$ peuvent être modélisées à partir de l'information auxiliaire $\mathbf{x}_k$, c'est-à-dire qu'il existe une fonction f_j telle que $c_{jk} = f_j(\mathbf{x}_k) + \alpha_{jk}$, $j = 1, ..., q$. Le modèle de superpopulation peut alors s'écrire de la façon suivante

$$\xi : \quad Y_k(t) \quad = \quad \mu(t) + \sum_{j=1}^{q} f_j(\mathbf{x}_k) v_j(t) + \epsilon_k(t) \tag{A.2}$$

avec $\epsilon_k(t) = \sum_{j=1}^{q} \alpha_{jk} v_j(t) + \eta_k(t)$.

Dans la réalité, Y_k n'est connu que pour $k \in s$. Dans un premier temps, il faut estimer l'opérateur de covariance Γ pour estimer les fonctions propres v_j associées aux q plus grandes valeurs propres. Une estimation de v_j est donnée par $\tilde{v}_j$, fonction propre de la matrice de covariance de Horvitz-Thompson :

$$\hat{\Gamma}(r, t) = \frac{1}{N} \sum_{s} \frac{(Y_k(t) - \hat{\mu}(t))(Y_k(r) - \hat{\mu}(r))}{\pi_k}$$

A la fin de cette étape, nous avons, pour tout $k \in s$,

$$\tilde{Y}_k(t) = \hat{\mu}(t) + \sum_{j=1}^{q} \tilde{c}_{jk}\tilde{v}_j(t)$$

où $\tilde{c}_{jk} =< Y_k - \hat{\mu}, \tilde{v}_j(t) >$.

Les $\tilde{c}_{jk}$ n'étant connus que pour $k \in s$, nous les modélisons à l'aide de l'information auxiliaire disponible. On obtient ainsi une estimation pour tout $k \in U$

$$\hat{\tilde{c}}_{jk} = \hat{f}_j(\mathbf{x}_k) = \arg\min_{f_j} \sum_{k\in s} \frac{1}{\pi_k} \left(\tilde{c}_{jk} - f_j(\mathbf{x}_k)\right)^2, \quad \forall k \in U \text{ et } j = 1, ..., q.$$

Finalement, l'estimateur assisté par le modèle basé sur l'ACP fonctionnelle est donné par

$$\hat{\mu}_{MA}(t) = \frac{1}{N} \sum_{k\in U} \hat{Y}_k(t) - \frac{1}{N} \sum_{k\in s} \frac{(\hat{Y}_k(t) - Y_k(t))}{\pi_k}$$

avec $\hat{Y}_k(t) = \hat{\mu}(t) + \sum_{j=1}^{q} \hat{f}_j(\mathbf{x}_k)\tilde{v}_j(t)$, $t \in [0, T]$.

Remarque A.1. *Cette approche permet de séparer l'effet temporel de l'effet des covariables. De plus, lorsqu'on peut obtenir une bonne approximation dans un espace de petite dimension, elle peut prendre en compte, de manière assez aisée, des effets non linéaires des covariables (cf. Chiou* et al. *(2003)).*

A.2 Estimation de la courbe moyenne de consommation des particuliers

Il est bien connu qu'il existe un lien entre la consommation électrique et la température extérieure au moins pour les clients dits "thermosensibles". Ce lien est bien visible sur la courbe de consommation annuelle des particuliers qui se chauffent au chauffage électrique. Leur consommation est élevée pendant les périodes froides et diminue quand la température extérieure augmente. Dans cette section, nous allons regarder si la prise en compte de ce lien permet d'améliorer ou non l'estimation de la courbe moyenne de consommation des particuliers. La température n'étant connue qu'au moment de l'échantillonnage, nous ne nous sommes intéressés qu'aux méthodes d'estimation avec l'approche modèle et plus particulièrement à l'estimateur basé sur la régression fonctionnelle et celui basé sur l'ACP. Dans un premier temps, nous allons utiliser la population U pour déterminer une estimation de la fonction f de notre modèle de superpopulation

$$\xi : \quad Y_k(t) = f(\mathbf{x}_k, t) + \epsilon_{kt}, \qquad k \in U \text{ et } t \in [0, T]$$

où les erreurs ϵ_{kt} sont des variables aléatoires, d'espérance nulle et de variance $v(\mathbf{x}_k) = v_k$. Dans un deuxième temps, nous comparerons différentes stratégies de sondage.

Nous disposons d'une population U composée de $N = 2272$ courbes de consommation électrique de particuliers mesurées toutes les 10 minutes pendant une journée de demi-saison, notée J. L'ensemble des individus de notre population possède un contrat de type "Heure Creuse-Heure Pleine". Nous disposons de 4 variables auxiliaires :

– la consommation moyenne de la semaine précédente.
– la température moyenne de la semaine précédente.
– la température extérieure relevée toutes les 3 heures les jours J, $J-1$ et $J-2$.
– l'indicatrice Heure Creuse pour chaque pas 10 minutes du jour J.

A chaque individu k, nous attribuons la température relevée à la station météorologique la plus proche. Plusieurs individus auront ainsi la même température.

A.2.1 Recherche du modèle de superpopulation

Avant de mettre en place un plan de sondage, nous allons modéliser, à partir de l'ensemble des individus de la population, la fonction f dans le modèle de superpopulation ξ.

Modélisation à l'aide de l'ACP fonctionnelle

Dans un premier temps, nous avons regardé la modélisation basée sur l'ACP fonctionnelle. Cette méthode permet de prendre en compte le lien entre les instants lors de la modélisation. Nous constatons que pour garder un maximum d'inertie (au moins 70%), il va falloir garder au moins 13 axes (cf Figure A.1) et chercher à modéliser autant de composantes à l'aide de l'information auxiliaire disponible.

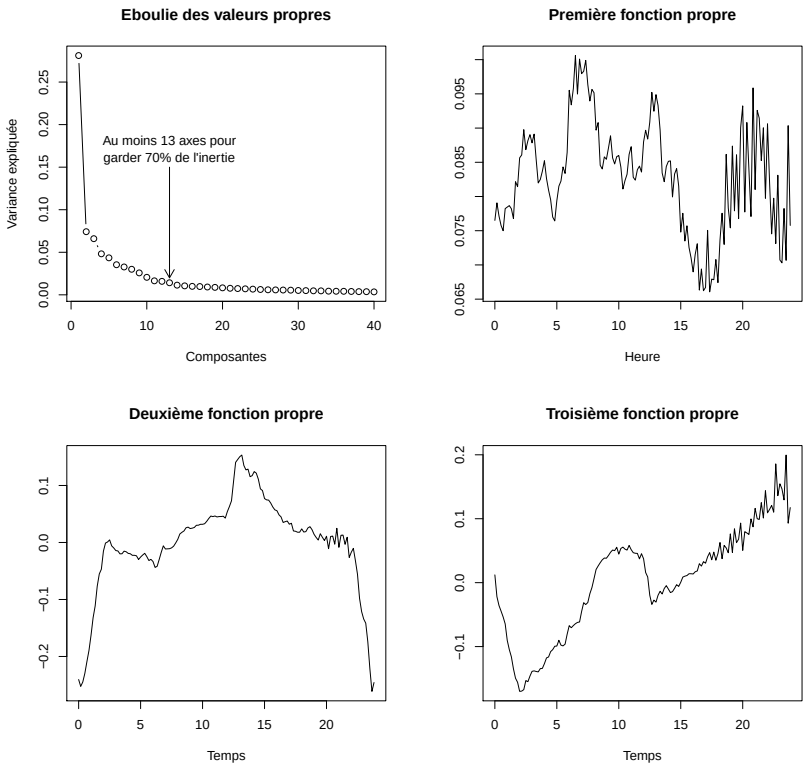
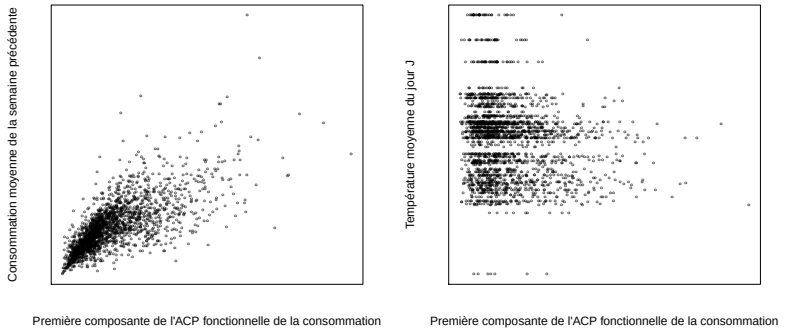


FIGURE A.1 – Sortie de l'ACP fonctionnelle du jour J

Remarque A.2. *Le comportement des particuliers est beaucoup plus hétérogène que celui des entreprises où les deux premières composantes permettaient de capturer la quasi totalité de l'inertie (cf. Cardot* et al. *(2010b)).*

Régression de la première composante

Nous avons cherché la meilleure régression linéaire pour modéliser la première composante (cf Figure A.2). Les variables les plus significatives sont la consommation moyenne de la semaine précédente et la température moyenne des jours $J-1$ et J (R^2 ajusté : 0.54). La température n'apporte pas grand chose, l'essentiel de la part expliquée provient de la consommation antérieure (le R^2 ajusté augmente seulement de 0.01).



(A) (B)

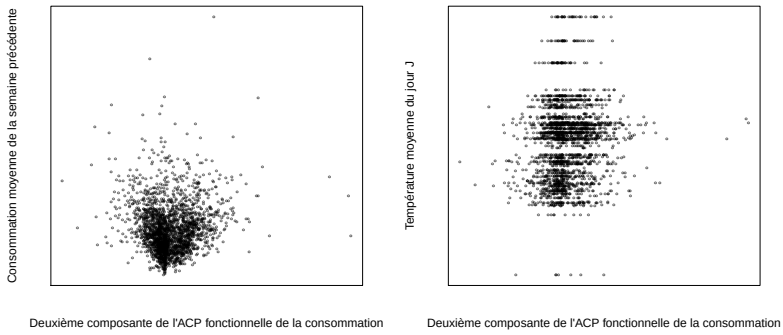
FIGURE A.2 – Représentation graphique entre la première composante et (a) la consommation moyenne de la semaine précédente (b) la température moyenne du jour J.

Modélisation de la deuxième composante

Nous avons représenté la deuxième composante en fonction de la température et de la consommation de la semaine précédente (cf. Figure A.3). Aucune forme ne se dégage de ces nuages de points (même avec des passages au log, à l'exponentielle ou à la puissance). Ces 2 variables ne permettent pas d'expliquer cette composante.

Conclusion

Dans notre cas, il faut conserver énormément d'axes pour garder un maximum d'inertie. Il va donc falloir trouver les variables auxiliaires qui nous permettent de les modéliser. Pour l'instant, nous pouvons modéliser uniquement la première composante et nous ne disposons pas d'assez de variables pour construire un bon modèle. La température n'intervient quasiment pas dans la modélisation. Nous avons testé différentes façons de la faire intervenir (lissage de la température, ACP fonctionnelle sur 3 jours, interaction entre les variables températures) et nous avons constaté que c'est encore la température moyenne de la journée qui intervient le plus. Pour conclure, avec le peu de données dont nous disposons, il nous est impossible d'utiliser le modèle basé sur l'ACP fonctionnelle et la température n'intervient que très légèrement.



(A) (B)

FIGURE A.3 – Représentation graphique entre la deuxième composante et (a) la consommation moyenne de la semaine précédente (b) la température moyenne du jour J.

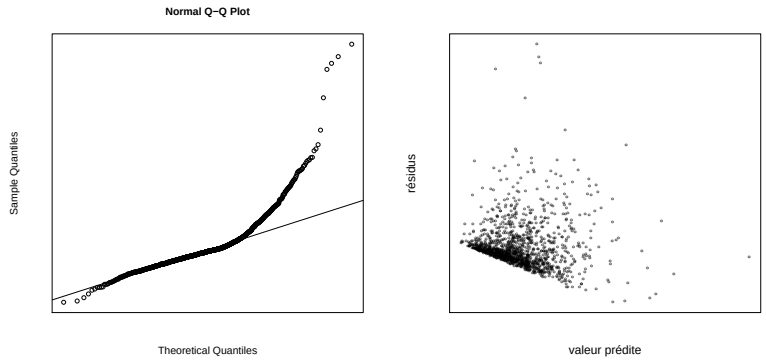
Modélisation instant par instant

Voyant les résultats peu concluants avec l'ACP fonctionnelle, nous nous sommes intéressés à la modélisation instant par instant pour rechercher d'éventuelles différences de modélisation entre les instants de cette journée de mi-saison (réaction différente à la température, l'influence du tarif Heure Creuse - Heure Pleine). Pour voir l'influence de la température dans l'estimation, nous avons décidé de faire une classification à partir de leur historique mensuel de consommation (classification de WARD réalisée à partir des coordonnées des individus sur les axes de l'AFC). Nous avons retenu 2 classes : les non-thermosensibles (N_{NT} = 951) et les thermosensibles (N_T = 1321)

Les thermosensibles

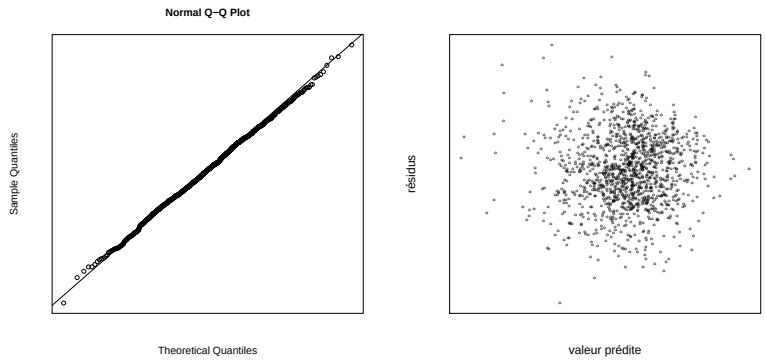
Dans un premier temps, nous avons étudié l'instant t_{max} où la consommation est la plus corrélée avec la température extérieure moyenne de la journée (corrélation maximum = -0.238). En faisant, la régression consommation $Y(t_{max})$ en fonction de la température moyenne, de la consommation antérieure et de la variable Heure Creuse, nous constatons la présence de valeurs extrêmes (cf. Figure A.4). Afin de réduire leur influence lors de la modélisation, nous avons transformé la variable consommation par passage au log (cf. Figure A.5). La température n'apporte pas grand chose, le R^2 ajusté augmente que de 0.003 (R^2 ajusté : 0.95). Les résultats sont à prendre avec précaution, nous ne faisons pas une régression directement sur la consommation mais sur une transformation de celle-ci. Il faudra éventuellement envisager un modèle non-paramétrique basé sur la variable $Z = log(Y)$ pour améliorer la précision.

Dans un deuxième temps, on s'est intéressé à l'instant t_{min} le moins corrélé avec la consommation (corrélation minimum = -0.0381). Lorsqu'on réalise la régression de la variable $Z = log(Y(t_{min}))$ en fonction de la température moyenne du jour J et du log de la consommation antérieure, deux groupes se dégagent du nuage de points valeurs



(A) Normalité (B) Homoscédasticité

FIGURE A.4 – Thermosensibles : Régression de la consommation à l'instant où la température est la plus corrélée avec la consommation



(A) Normalité (B) Homoscédasticité

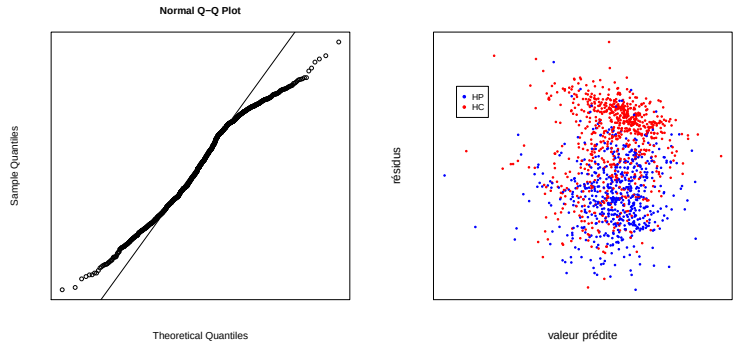
FIGURE A.5 – Thermosensibles : Régression du log de la consommation à l'instant où la température est la plus corrélée avec la consommation

prédites en fonction des résidus (cf. Figure A.6). L'introduction de la variable Heure Creuse nous a permis de les identifier et d'augmenter légèrement le R^2 ajusté, égal à 0.95 (cf. Figure A.7). Pour améliorer ce modèle, il faudrait éventuellement introduire une variable qui nous indique depuis quand l'individu est en Heure Creuse. Cette nouvelle variable nous permettrait de prendre en compte l'appel de puissance dû au déclenchement du chauffe eau, à la mise en route de la machine à laver, etc.

Les non-thermosensibles

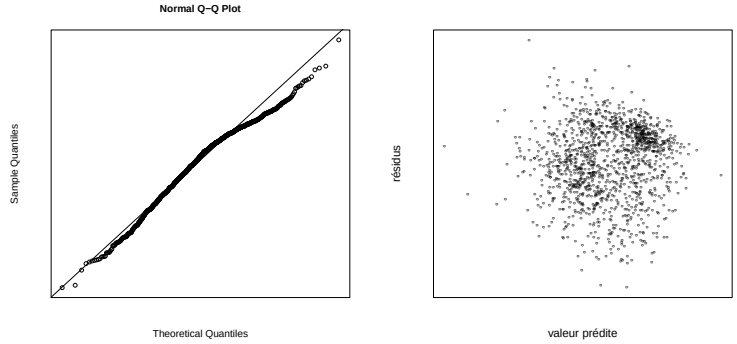
Comme nous nous y attendions, pour les clients non-thermosensibles la température n'intervient pas dans la modélisation de la consommation à l'instant t (cf. Figure A.8). Leur consommation est très liée à celle de la semaine précédente.

Nous avons alors considéré l'estimateur assisté par un modèle sur la variable d'intérêt transformée. Notons $Y_k(t)$ la consommation de l'individu k à l'instant t, $conso_prec_k$ la consommation moyenne de la semaine précédente de l'individu k, $mean_temp_k$ la



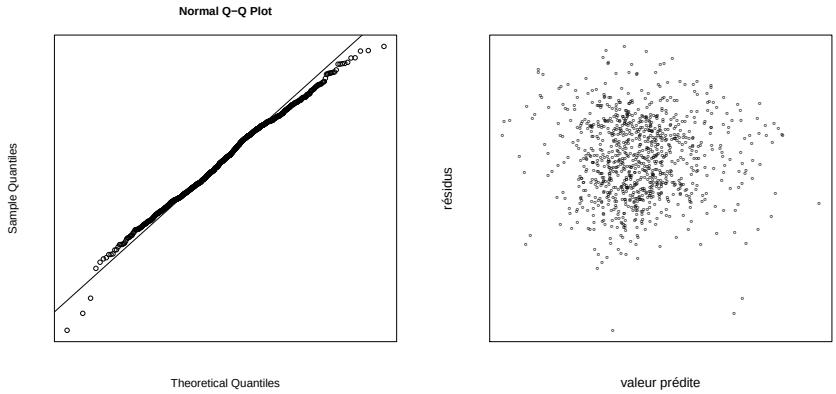
(A) Normalité (B) Homoscédasticité

FIGURE A.6 – Thermosensibles : Régression du log de la consommation à l'instant où la température est la moins corrélée avec la consommation (sans Heure Creuse)



(A) Normalité (B) Homoscédasticité

FIGURE A.7 – Thermosensibles : Régression du log de la consommation à l'instant où la température est la moins corrélée avec la consommation (avec Heure Creuse)



(A) Normalité (B) Homoscédasticité

FIGURE A.8 – Non-Thermosensibles : Régression à l'instant où la température est la plus corrélée avec la consommation

température moyenne de la journée d'étude de l'individu k et $HC_k(t)$ l'indicatrice Heure Creuse de l'individu k à l'instant t.

On suppose que $\forall k \in U$ et $t \in [0,T]$,

$$log(Y_k(t)) = \beta_1(t)log(conso_prec_k) + \beta_2(t)HC_k(t) + \beta_3(t)mean_temp_k + \eta_{k,t} \quad (A.3)$$

où η_{kt} sont indépendants, $\mathbb{E}_\xi(\eta_k) = 0$, de fonction de covariance $\mathrm{Cov}_\xi(\eta_{kt}, \eta_{kr}) = \Gamma(t,r)$ si $k = l$ et 0 sinon, pour $(t,r) \in [0,T] \times [0,T]$.

1. On tire l'échantillon s à l'aide d'un sondage aléatoire simple sans remise.
2. On estime le vecteur $\beta(t) = (\beta_1(t), \beta_2(t), \beta_3(t))'$, pour tout $t \in [0,T]$, à l'aide de l'échantillon s.

$$\hat{\beta}(t) = \left(\sum_s \frac{\mathbf{x}_k(t)\mathbf{x}_k(t)'}{\pi_k}\right)^{-1} \sum_s \frac{\mathbf{x}_k(t)log(Y_k(t))}{\pi_k}, \quad (A.4)$$

   où $\mathbf{x}_k(t) = (log(conso_prec_k), HC_k(t), mean_temp_k)'$
3. On calcule, pour tout $t \in [0,T]$, $\hat{Y}_k(t) = e^{\mathbf{x}_k'(t)\hat{\beta}(t)}$.
4. La courbe moyenne μ est estimée par

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

A.2.2 Mise en place du plan de sondage

Dans cette section, nous allons comparer différentes stratégies (plan de sondage avec estimateur) qui prennent en compte la température pour estimer la courbe moyenne du jour J.

8 stratégies ont été retenues :

- Stratégie 1 : Sondage aléatoire simple sans remise (SRSWOR) avec l'estimateur de Horvitz-Thompson.
- Stratégie 2 : Sondage à probabilités inégales sans remise avec $\pi_k = n\frac{conso_prec_k}{\sum_U conso_prec_k}$ et avec l'estimateur de Horvitz-Thompson.
- Stratégie 3 : Sondage SRSWOR avec l'estimateur assisté par le modèle de régression linéaire suivant

$$Y_k(t) = \beta_1(t)conso_prec_k + \beta_2(t)HC_k(t) + \beta_3(t)mean_temp_k + \eta_{kt}$$

- Stratégie 4 : Sondage SRSWOR avec l'estimateur assisté par le modèle de régression linéaire suivant

$$Y_k(t) = \beta_1(t)conso_prec_k + \beta_2(t)HC_k(t) + \beta_3(t)temp_spline_k(t) + \eta_{kt}$$

- Stratégie 5 : Sondage SRSWOR avec l'estimateur assisté par le modèle de régression linéaire suivant

$$Y_k(t) = \beta_1(t)conso_prec_k + \beta_2(t)HC_k(t) + \eta_{kt}$$

– Stratégie 6 : Sondage SRSWOR avec l'estimateur assisté par le modèle sur la variable d'intérêt transformée suivant

$$log(Y_k(t)) = \beta_1(t)log(conso_prec_k) + \beta_2(t)HC_k(t) + \beta_3(t)mean_temp_k + \eta_{kt}$$

– Stratégie 7 : Sondage SRSWOR avec l'estimateur assisté par le modèle de régression linéaire suivant

$$Y_k(t) = \beta_0(t) + \beta_1(t)mean_temp_k + \eta_{kt}$$

– Stratégie 8 : Sondage SRSWOR avec l'estimateur obtenu par un calage sur la base de B-spline de conso_prec (d'ordre deux avec 3 nœuds)

$$\hat{\mu}(t) = \frac{1}{N}\sum_s \omega_{ks}^b Y_k(t)$$

avec $w_{ks}^b = d_k\left[1 - q_k\mathbf{b}'(z_k)\hat{\mathbf{T}}_s^{-1}(\hat{t}_{\mathbf{b},d} - t_{\mathbf{b}})\right]$, $\mathbf{b}(z_k) = (B_1(z_k), \ldots, B_5(z_k))'$, $\hat{\mathbf{T}}_s = \sum_s d_k q_k \mathbf{b}(z_k)\mathbf{b}'(z_k)$, $\hat{t}_{\mathbf{b},d} = \sum_s d_k \mathbf{b}(z_k)$, $t_{\mathbf{b}} = \sum_U \mathbf{b}(z_k)$ et $q_k = 1$.

La variable $temp_spline_k(t)$ a été obtenue en appliquant un lissage B-spline sur les températures tri-horaire du jour J. Dans la stratégie 2, nous avons tiré notre échantillon à l'aide de l'algorithme du cube équilibré sur les π_k. Le sondage πps systématique a également été testé mais celui-ci donne des résultats moins bons.

Nous avons estimé séparément la courbe moyenne des clients thermosensibles (noté T) et celle des clients non-thermosensibles (noté NT). La taille de l'échantillon est fixée à $n_T = 150$ (resp. $n_{NT} = 100$) pour les clients thermosensibles (resp. les non-thermosensibles). Dans les stratégies 3, 4, 6 et 7, nous avons décidé d'inclure la variable température pour tous les instants même si pour certains nous savons qu'elle n'est pas significative. Nous voulions garder un modèle identique entre les instants.

Pour évaluer la précision de nos estimateurs $\hat{\mu}$, nous avons réalisé 10000 simulations et nous avons comparé l'erreur moyenne

$$R(\hat{\mu}) = \int_0^T |\hat{\mu}(t) - \mu(t)|dt$$

Le passage au log (stratégie 6) ne permet pas d'améliorer la précision de l'estimation de la courbe moyenne. Pour des raisons de simplicité, il faut mieux utiliser le modèle basé sur la régression.

L'utilisation de l'information auxiliaire a permis d'améliorer la précision des estimateurs des stratégies 2 à 6. Lorsqu'on compare les stratégies 3 à 5, nous constatons que l'utilisation de la variable température (stratégie 3 et 4) permet d'améliorer très légèrement la précision de l'estimation des clients thermosensibles. De plus, il n'est pas nécessaire d'utiliser la variable température sous forme fonctionnelle (stratégie 4), la température moyenne de la journée (stratégie 3) est suffisante. Pour les clients non thermosensibles, la température n'apporte aucune amélioration. Si on souhaite travailler

	Moyenne	1e quantile	médiane	3e quantile
Stratégie 1	12.48908	10.43878	11.86091	13.81739
Stratégie 2	11.15794	9.76768	10.86895	12.23394
Stratégie 3	11.18751	9.826591	10.929534	12.197746
Stratégie 4	11.19351	9.790156	10.893169	12.304361
Stratégie 5	11.18827	9.804525	10.898856	12.264005
Stratégie 6	11.25619	9.89151	10.98851	12.32621
Stratégie 7	12.37778	10.42652	11.78828	13.64685
Stratégie 8	12.22762	10.422026	11.683725	13.319986

Table A.1 – Erreur moyenne des clients thermosensibles

	Moyenne	1e quantile	médiane	3e quantile
Stratégie 1	10.33943	8.93304	9.97682	11.31109
Stratégie 2	7.26167	6.53261	7.13508	7.85494
Stratégie 3	9.34171	8.38999	9.20619	10.12591
Stratégie 4	9.34158	8.38346	9.21655	10.13778
Stratégie 5	9.31090	8.37953	9.17886	10.08273
Stratégie 6	9.48277	8.50125	9.33075	10.30681
Stratégie 7	10.3718	8.97320	10.03747	11.36665
Stratégie 8	10.49556	8.88109	9.89708	11.29610

Table A.2 – Erreur moyenne des clients non-thermosensibles

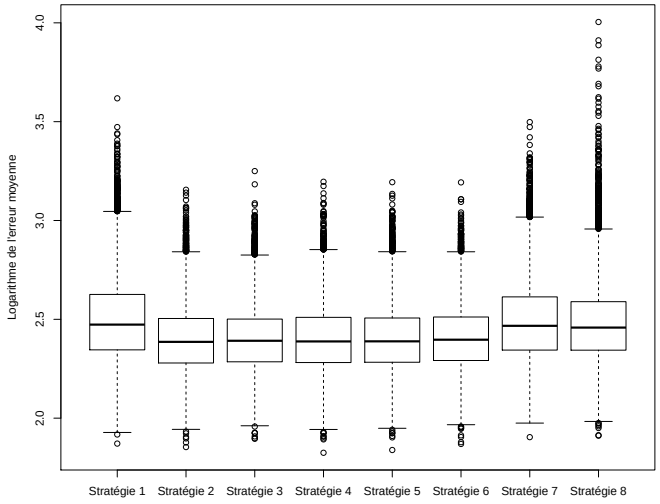


FIGURE A.9 – Comparaison des erreurs moyennes d'estimation des clients thermosensibles

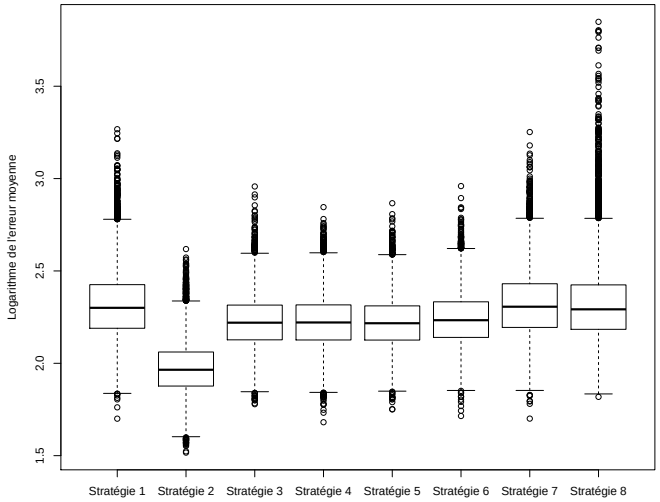


FIGURE A.10 – Comparaison des erreurs moyennes d'estimation des clients non-thermosensibles

avec l'estimateur assisté par le modèle de régression linéaire, il vaut mieux utiliser la stratégie 5 pour les non-thermosensibles et la stratégie 3 pour les thermosensibles.

Avec le modèle basé uniquement sur la température (stratégie 7), on considère que tous les individus qui ont la même température auront la même consommation prédite. Or nous savons très bien que deux individus soumis aux mêmes températures extérieures auront forcément une consommation différente (isolation et installation

électrique différentes). Il nous parait donc normal que ce modèle ne permette pas d'améliorer l'estimation.

Le calage sur les fonctions B-splines de la consommation antérieure (stratégie 8) ne permet pas d'améliorer la précision de l'estimateur. Ce résultat peut être dû au fait que nous travaillons sur une période de mi-saison : il y a un changement de comportement par rapport à la semaine précédente.

Pour les clients non-thermosensibles, le sondage à probabilités inégales sans remise améliore significativement la précision de l'estimateur. Par contre, pour les clients thermosensibles, le gain est moins élevé. Il est comparable à celui obtenu avec les stratégies faisant appel à l'estimateur assisté par le modèle de régression linéaire. Le plan à probabilités inégales donne de meilleurs résultats pour les clients non-thermosensibles car la corrélation entre la consommation de la journée d'étude et la consommation de la semaine précédente est plus élevée que sur les clients thermosensibles.

Les simulations effectuées sur cette journée de mi-saison nous montrent que le modèle basé sur la régression, mis en place pour prendre en compte la température extérieure, ne permet pas d'améliorer la précision et que les meilleurs résultats sont obtenus à l'aide du plan à probabilités inégales. Toutefois, il ne faut pas complètement rejeter le modèle proposé. Dans notre étude, nous avons constaté que le modèle est équivalent en termes de précision au sondage à probabilités inégales pour les clients thermosensibles. Si nous disposons de nouvelles variables auxiliaires, l'approche modèle pourrait donner de meilleurs résultats que le sondage à probabilités inégales pour les journées de mi-saison ou pour les journées où il y a une chute brutale des températures extérieures. Pour les journées d'hiver ou d'été, le changement de comportement individuel est moins marqué. La consommation de la journée est très corrélée à la consommation de la semaine précédente pour nos deux populations. Il est donc fort probable que, pour ces journées, un sondage à probabilités inégales suffise.

A.3 Conclusion des travaux

Dans cette étude, nous avons cherché à prendre en compte la variable température extérieure dans un plan de sondage. Nos recherches nous ont amené à construire un modèle d'estimation basé sur la régression. Nous avons essayé d'introduire dans ce modèle la température sous différentes formes (température observée à chaque instant, température moyenne de la veille, lissage, puissance de la température moyenne de la journée) et nous avons constaté que c'est la température moyenne de la journée qui est la plus significative. L'ajout de cette variable dans notre modèle n'a permis d'améliorer que très légèrement l'estimation pour les clients thermosensibles. Ce très faible apport peut s'expliquer par le fait que la température est en quelque sorte déjà prise en compte dans la variable consommation de la semaine antérieure ou par son manque de variabilité. En effet, tous les individus qui sont reliés à la même station météorologique auront la même température, or, dans la réalité, la température des individus peut

être différente de celle de la station météorologique. La variable température est donc entachée d'une erreur d'observation.

Dans un deuxième temps, nous avons comparé l'estimateur basé sur le modèle (stratégie 3) au sondage aléatoire simple (stratégie 1) et au sondage à probabilités inégales (stratégie 2). Sur notre jeu de données, la stratégie 3 nous permet d'avoir une meilleure estimation que le sondage aléatoire simple sans remise avec Horvitz-Thompson (en moyenne l'erreur est diminuée de 11.65% pour les thermosensibles et de 11.05% pour les non-thermosensibles). Par contre, nous constatons que cette stratégie est moins bonne pour les clients non-thermosensibles que le sondage à probabilités inégales et qu'elle est équivalente pour les clients thermosensibles (en moyenne l'erreur de la stratégie 2 par rapport à celle de la stratégie 1 est diminuée de 11.92% pour les thermosensibles et de 42.38% pour les non-thermosensibles). Les divers résultats nous ont montré que c'est essentiellement la consommation de la semaine précédente et la variable Heure Creuse qui expliquent la consommation de la journée d'étude et que la température apporte très peu d'amélioration. Nous avons également constaté qu'en utilisant un modèle qui prend en compte uniquement la température pour estimer la consommation nous obtenons de plus mauvais résultats qu'un sondage aléatoire sans remise avec Horvitz-Thompson. Donc sans apport de nouvelles variables auxiliaires, il semble préférable d'utiliser un plan à probabilités inégales. De plus, pour ce plan l'estimateur est sans biais et des approximations de la variance sont connues.

D'autres pistes de modèle ont été envisagées (modèle basé sur l'ACP fonctionnelle et modèle utilisant la décomposition en ondelettes) mais nous ne disposons pas d'assez de variables auxiliaires pour pouvoir les appliquer. Nous avons également envisagé d'effectuer une post-stratification sur la variable station météo mais nous disposions de trop peu d'individus par station pour pouvoir l'appliquer.

Pour confirmer nos conclusions, il faudrait tester notre modèle pour une journée de mi-saison et une journée d'hiver sur un jeu de données plus grand et avec éventuellement de nouvelles variables auxiliaires. Pour faire ce test, il sera nécessaire de connaitre l'historique de consommation mensuelle des individus pour pouvoir faire la classification thermosensible.

Bibliographie

Adler, R. J. et Taylor, J. E. (2007). *Random fields and geometry.* Springer-Verlag, New York.

Ardilly, P. (2006). *Les techniques de sondage.* Editions TECHNIP.

Beaumont, J. F. et Rivest, L. P. (2009). Dealing with outliers in survey data. *In* Pfeffermann, D. et Rao, C., éditeurs : *Handbook of statistics*, volume 29A, pages 247–279. Elsevier.

Berger, Y. G. (1998a). Rate of convergence for asymptotic variance of the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 74:149–168.

Berger, Y. G. (1998b). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 67:209–226.

Berger, Y. G. et Rao, J. (2006). Adjusted jackknife for imputation under unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, B68:531–547.

Bickel, P. J. et Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The annals of statistics*, 12:470–482.

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

Billingsley, P. (1968). *Convergence of Probability Measures.* John Wiley and Sons.

Boistard, H., Lopuhaä, H. P. et Ruiz-Gazen, A. (2012). Approximation of rejective sampling inclusion probabilites and application to higher order correlation. Rapport technique, Arxiv 1207.5654.

Bondesson, L., Traat, I. et Lundqvist, A. (2006). Pareto sampling versus Sampford and conditional Poisson sampling. *Scand. J. Statist.*, 33(4):699–720.

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

Bosq, D. (2000). *Linear Processes in Function Spaces : Theory and Applications*, volume 149 de *Lecture notes in Statistics.* Springer-Verlag, New York.

Breidt, F. J., Claeskens, G. et Opsomer, J. D. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92:831–846.

Breidt, F. J. et Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *The annals of statistics*, 28(4):1023–1053.

Breidt, F. J. et Opsomer, J. D. (2009). Nonparametric and semiparametric estimation in complex surveys. *In* Pfeffermann, D. et Rao, C., éditeurs : *Handbook of statistics - Sample surveys : Inference and analysis*, volume 29B, pages 103–119. Elsevier.

Brewer, K. et Hanif, M. (1983). *Sampling with unequal probabilities.* Springer-Verlag, New York.

Brick, M. J. et Montequila, J. M. (2009). Nonresponse and weighting. *In* Pfeffermann, D. et Rao, C., éditeurs : *Handbook of statistics*, volume 29A, pages 163–175. Elsevier.

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

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

Cardot, H., Degras, D. et Josserand, E. (2012). Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data. *to appear in Bernoulli.*

Cardot, H., Dessertaine, A. et Josserand, E. (2010b). Semiparametric models with functional responses in a model assisted survey sampling setting. *In* Lechevallier, Y. et Saporta, G., éditeurs : *Compstat 2010*, pages 411–420. Physica-Verlag, Springer.

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

Cassel, C. M., Särndal, C. E. et Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63:615–620.

Chaouch, M. et Goga, C. (2012). Using complex surveys to estimate the l_1-median of a functional variable : application to electricity load curves. *international Statistical Review*, 80(1):40–59.

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

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

Chen, X.-H., Dempster, A. P. et Liu, J. S. (1994). Weighting finite population sampling to maximise entropy. *Biometrika*, 81:457–469.

Chiky, R. (2009). *Résumé de flux de données distribuées.* Thèse de doctorat, l'Ecole Nationale Supérieure des Télécommunications.

Chiou, J. M., Müller, H. G. et Wang, J. L. (2003). Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society : Series B*, 65(2):405–423.

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

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

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

Davison, A. et Hinkley, D. (1997). *Cambridge University Press, Cambridge.* Bootstrap Methods and Their Application.

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

Degras, D. (2012). Rotation sampling for functional data. http ://arxiv.org/abs/1204.4494.

Demnati, A. et Rao, J. N. K. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology*, 30:17–34.

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

Deville, J. C. (1993). Estimation de la variance pour les enquêtes à deux phases. Manuscript. INSEE, Paris.

Deville, J. C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25(2):219–230.

Deville, J. C. (2000). Note sur l'algorithme de Chen, Dempster et Liu. Technical report, CREST-ENSAI.

Deville, J. C. et Särndal, C. E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, 87:376–382.

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

Deville, J. C. et Tillé, Y. (2005). Variance approximation under balanced sampling. *J. Statist. Plann. Inference*, 128:569–591.

Dierckx, P. (1993). *Curve and surface fitting with splines.* Oxford, Clarendon Press.

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

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

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

Ferraty, F. et Vieu, P. (2006). *Nonparametric functional data analysis.* Springer series in statistics.

Fuller, W. A. (2009a). *Sampling statistics.* John Wiley and sons.

Fuller, W. A. (2009b). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4):933–944.

Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliaire : une approche nonparamétrique par splines de régression. *The Canadian Journal of Statistics*, 33(2):1–18.

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

Guillas, S. (2001). Rates of convergence of autocorrelation estimates for autoregressive Hilbertian processes. *Statist. and Probability Letters*, 55:281–291.

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

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

Hájek, J. (1981). *Sampling from a finite population.* Statistics : Textbooks and Monographs. Marcel Dekker, New York.

Hall, P., Müller, H. G. et Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The annals of statistics*, 24:1493–1517.

Hansen, M. H. et Hurvitz, W. N. (1943). on the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14:333–362.

Haziza, D. (2009). Imputation and inference in the presence of missing data. *In* Pfeffermann, D. et Rao, C., éditeurs : *Handbook of statistics*, volume 29A, pages 215–246. Elsevier.

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

HORVITZ, D. et 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. et FULLER, W. A. (1982). Survey design under the regression superpopulation model. *J. AM. Statist. Ass.*, 77:49–61.

KEMPERMAN, J. H. B. (1969). On the optimum rate of transmitting information. *Ann. Math. Statist.*, 40:2156–2177.

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

LUNDSTRÖM, S. et SÄRNDAL, C. E. (1999). Calibration as a standard method for treatement of nonresponse. *Journal of Official Statistics*, 15(2):305–327.

MADOW, W. G. (1949). On the theory of systematic sampling, II. *Annals of Mathematical sampling*, 20:333–354.

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

MATEI, A. et TILLÉ, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21(4):543–570.

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.

PITT, L. D. et TRAN, L. T. (1979). Local sample path properties of Gaussian fields. *Annals of Probability*, 7:477–493.

QUENOUILLE, M. (1949). Approximate tests of correlation in times-series. *Journal of the Royal Statistical Society*, B11:68–84.

RAMSAY, J. O. et SILVERMAN, B. W. (2005). *Functional data analysis*. Springer series in statistics.

RAO, J. et WU, C. (1985). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83:231–241.

RAO, J. et WU, C. (1988). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18:209–217.

ROBINSON, P. M. et SÄRNDAL, C. E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya : The Indian Journal of Statistics*, 45:233–243.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57:377–387.

Särndal, C. E. (1980). On π inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67:639–50.

Särndal, C. E. et Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* New York : John Wiley and Sons.

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

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

Shao, J. et Tu, D. (1995). *The Jakknife and The Bootstrap.* Springer, New-York.

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

Staniswalis, J. G. et Lee, J. J. (1998). Nonparametric regression analysis of longitudinal data. *J. Amer. Statist. Assoc.*, 93:1403–1418.

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

Tillé, Y. (2006). *Sampling algorithms.* Springer Series in Statistics. Springer, New York.

Tillé, Y. (2011). Ten years of balanced sampling with the cube method : an appraisal. *Survey Methodology*, 37:215–226.

Tuckey, J. (1958). Biais and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29:614.

Valliant, R., Dorfman, A. H. et Royall, R. M. (2000). *Finite Population Sampling and Inference : A Prediction Approach.* John Wiley and Sons, New York.

Van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

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

Víšek, J. Á. (1979). Asymptotic distribution of simple estimate for rejective, Sampford and successive sampling. *In Contributions to statistics*, pages 263–275. Reidel, Dordrecht.

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

Zhou, S., Shen, X. et Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *The annals of statistics*, 26(5):1760–1782.