UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

# Inférence basée sur le plan pour l'estimation de petits domaines
# Design-based inference for small area estimation

**Thèse présentée à la Faculté des Sciences Economiques**

Ecole Doctorale **OMI**

Université Paris-Est

pour l'obtention du grade de docteur en Statistique

par

**Toky Randrianasolo**

Acceptée par le jury de thèse :

**Prof. Yves Tillé,** Université de Neuchâtel, directeur de thèse

**Prof. Jean-Loup Madre,** Université Paris-Est, IFSTTAR-DEST, directeur de thèse

**Prof. David Haziza,** Université de Montréal

**Prof. Stephan Morgenthaler,** Ecole Polytechnique Fédérale de Lausanne

**Dr. Jimmy Armoogum**, Université Paris-Est, IFSTTAR-DEST

Soutenue le 18 novembre 2013

**Faculté des Sciences Economiques**
**Avenue du 1er-Mars 26**
**CH-2000 Neuchâtel**
www.unine.ch/seco

LETTRES ET SCIENCES HUMAINES
SCIENCES
DROIT
SCIENCES ECONOMIQUES
THEOLOGIE

Université de Neuchâtel
Institut de Statistique

Université Paris-Est
École Doctorale **OMI**

# Inférence basée sur le plan pour l'estimation de petits domaines

-

# Design-based inference for small area estimation

Thèse présentée pour l'obtention du grade de
Docteur en Statistique

par

Toky Randrianasolo

Acceptée sur proposition du jury composé de :

| | | |
|---|---|---|
| David Haziza | Université de Montréal | Président du jury & Rapporteur |
| Jean-Loup Madre | Université Paris-Est, Ifsttar-DEST | Directeur de thèse |
| Yves Tillé | Université de Neuchâtel | Directeur de thèse |
| Stephan Morgenthaler | École Polytechnique Fédérale de Lausanne | Rapporteur |
| Jimmy Armoogum | Université Paris-Est, Ifsttar-DEST | Examinateur |

Thèse soutenue le 18 novembre 2013.

UNIVERSITÉ DE NEUCHÂTEL

**IMPRIMATUR POUR LA THÈSE**

**(co-tutelle avec l'Université Paris-Est)**

Inférence basée sur le plan pour l'estimation de petits domaines

**Toky Randrianasolo**

---

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

La Faculté des sciences économiques,
sur le rapport des membres du jury

Prof. Yves Tillé (directeur de thèse, Université de Neuchâtel)
Prof. Jean-Loup Madre (directeur de thèse, Université Paris-Est)
Prof. David Haziza (Université de Montréal)
Prof. Stephan Morgenthaler (EPFL, Lausanne)
M. Jimmy Armoogum, chargé de recherche (Université de Paris-Est)

Autorise l'impression de la présente thèse.

Neuchâtel, le 17 décembre 2013

Le doyen

Jean-Marie Grether

■ Téléphone : +41 32 718 15 00 ■ Fax : +41 32 718 15 01 ■ E-mail : secretariat.seco@unine.ch ■ www.unine.ch

*"Un statisticien est un professionnel qui collecte diligemment des faits et des données, pour ensuite créer, avec attention, toutes sortes de confusions à leurs sujets."*
In *Winkler (2011)*.

*Tous les nombres premiers sont impairs sauf un.*
*Tous les nombres premiers sont impairs sauf deux.*

# Remerciements

Ce manuscrit de thèse est l'aboutissement de trois années de travaux de recherche dirigés par Yves Tillé et Jean-Loup Madre, et encadrés par Jimmy Armoogum. Jimmy, Yves, Jean-Loup, mes sincères remerciements pour vos précieux conseils.

Je tiens également à remercier Prof. David Haziza et Prof. Stephan Morgenthaler d'avoir accepté d'être les rapporteurs de cette thèse.

Je voudrais également remercier tous les collègues et anciens collègues du laboratoire DEST ainsi que ceux de l'Institut de Statistique de l'UniNe pour m'avoir offert un très bon cadre de travail. Un grand merci également à ceux qui ont bien voulu relire quelques parties de ce manuscrit.

Tout au long de ces trois dernières années, je remercie ma famille qui a su éviter de parler de l'avancée de ma thèse durant les repas. Aussi, mèrsi vilmools à mes amis pour leurs encouragements ainsi que pour les sorties récurrentes chez *Flam's* afin de retrouver un air du *pays*. Enfin et surtout, merci à ma pharmacognosienne préférée de m'avoir concocté toutes ces pâtisseries expérimentales.

Bry-sur-Marne, le 10 octobre 2013.

# Résumé

La forte demande de résultats à un niveau géographique fin, notamment à partir d'enquêtes nationales, a mis en évidence la fragilité des estimations sur petits domaines. Cette thèse propose d'y remédier avec des méthodes spécifiques basées sur le plan de sondage. Celles-ci reposent sur la construction de nouvelles pondérations pour chaque unité statistique. La première méthode consiste à optimiser le redressement du sous-échantillon d'une enquête inclus dans un domaine. La deuxième repose sur la construction de poids dépendant à la fois des unités statistiques et des domaines. Elle consiste à scinder les poids de sondage de l'estimateur global tout en respectant deux contraintes : 1/ la somme des estimations sur toute partition en domaines est égale à l'estimation globale ; 2/ le système de pondération pour un domaine particulier satisfait les propriétés de calage sur les variables auxiliaires connues pour le domaine. L'estimateur par scission ainsi obtenu se comporte de manière *quasi* analogue au célèbre estimateur BLUP (meilleur prédicteur linéaire sans biais). La troisième méthode propose une réécriture de l'estimateur BLUP sous la forme d'un estimateur linéaire homogène, en adoptant une approche basée sur le plan de sondage, bien que l'estimateur dépende d'un modèle. De nouveaux estimateurs BLUP modifiés sont obtenus. Leur précision, estimée par simulation avec application sur des données réelles, est assez proche de celle de l'estimateur BLUP standard. Les méthodes développées dans cette thèse sont ensuite appliquées à l'estimation d'indicateurs de la mobilité locale à partir de l'Enquête Nationale sur les Transports et les Déplacements 2007-2008. Lorsque la taille d'un domaine est faible dans l'échantillon, les estimations obtenues avec la première méthode perdent en précision, alors que la précision reste satisfaisante pour les deux autres méthodes.

**Mots-clés** sondage, estimation sur petits domaines, inférence basée sur le plan de sondage, poids.

# Abstract

The strong demand for results at a detailed geographic level, particularly from national surveys, has raised the problem of the fragility of estimates for small areas. This thesis addresses this issue with specific methods based on the sample design. These ones consist of building new weights for each statistical unit. The first method consists of optimizing the re-weighting of a sub-sample survey included in an area. The second one is based on the construction of weights that depend on the statistical units as well as the areas. It consists of splitting the sampling weights of the overall estimator while satisfying two constraints: 1/ the sum of the estimates on every partition into areas is equal to the overall estimate ; 2/ the system of weights for a given area satisfies calibration properties on known auxiliary variables at the level of the area. The split estimator thus obtained behaves almost similarly as the well-known BLUP (best linear unbiased predictor) estimator. The third method proposes a rewriting of the BLUP estimator, although model-based, in the form of a homogenous linear estimator from a design-based approach. New modified BLUP estimators are obtained. Their precision, estimated by simulation with an application to real data, is quite close to that of the standard BLUP estimator. Then, the methods developed in this thesis are applied to the estimation of local mobility indicators from the 2007-2008 French National Travel Survey. When the size of an area is small in the sample, the estimates obtained with the first method are not precise enough whereas the precision remains satisfactory for the two other methods.

**Keywords** survey sampling, small area estimation, design-based inference, weights.

# Table des matières

# Liste des figures

# Liste des tableaux

# Liste des algorithmes

# Introduction

Outre l'estimation de quantités (un total, une moyenne, etc.) au niveau de l'ensemble de la population, les données recueillies par un sondage peuvent être utilisées pour estimer une fonction d'intérêt pour une sous-population appelée *domaine*. Ceci découle de la forte demande toujours croissante de résultats à un niveau (généralement) géographique de plus en plus fin à partir d'enquêtes nationales. Le terme *domaine* peut toutefois référer à diverses variétés de sous-populations. Ces sous-populations peuvent par exemple représenter des zones géographiques comme une région, un département, un arrondissement, etc. Mais elles peuvent également représenter des groupes socio-démographiques obtenus éventuellement par croisement de plusieurs variables catégorielles. Un *domaine* peut donc être aussi bien géographique que catégoriel. Certaines littératures canadiennes font d'ailleurs la distinction : le terme *région* est employé si l'on se réfère à une zone géographique, le terme *domaine* est employé dans le cas d'une catégorie. Il en est de même dans les littératures anglo-saxones : le terme *area* dans le cas géographique et le terme *domain* le cas échéant du point de vue catégoriel. Cependant depuis quelques temps, par abus de langage peut-être, ces distinctions ont laissé place au terme *domaine* ou *area* en anglais, quelle que soit la caractéristique de la sous-population considérée. Il en sera de même dans la suite de ce manuscrit. Plus formellement donc, les *domaines* peuvent être simplement vus comme des sous-ensembles formant une partition de la population.

Les grandes enquêtes ont généralement pour but de fournir des estimations fiables au niveau national et à des niveaux (géographiques ou catégoriels) suffisamment agrégés, comme dans le cas des ZEAT[1]. La phase d'échantillonnage de ces grandes enquêtes ne prend pas forcément en

[1] Selon l'Insee, *"les zones d'études et d'aménagement du territoire ou* ZEAT *sont des subdivisions territoriales Françaises définies en 1967 par l'*Insee *en relation avec le Commissariat général au plan et la Délégation à l'aménagement du territoire et à l'action régionale. Huit super-régions de la France Métropolitaine sont alors créées : la région parisienne (Ile de France), le bassin parisien (Bourgogne, Centre, Champagne-Ardenne, Basse-Normandie, Haute-Normandie, Picardie), le Nord (Nord Pas-de-Calais), l'Est (Alsace, Franche-Comté, Lorraine), l'Ouest (Bretagne, Pays de la Loire, Poitou-Charentes), le Sud-Ouest (Aquitaine, Limousin, Midi-Pyrénées), le Centre-Est*

compte certains domaines au niveau desquels l'on pourrait ultérieurement être amené à fournir des estimations. Les tailles de ces domaines peuvent donc s'avérer faibles, ou très faibles, voire nulles, dans l'échantillon de l'enquête. On parlera alors de *petit* domaine. Plus formellement, un domaine est *petit* lorsqu'une estimation directe au niveau de celui-ci, i.e. une estimation construite uniquement avec l'information disponible au niveau de celui-ci, n'est pas suffisamment précise. La qualification de *petit* domaine est donc liée à la précision visée pour les estimations.

L'approche classique pour réaliser une estimation sur petits domaines consiste à recourir à un modèle. On utilise généralement un modèle linéaire mixte qui prend en compte les domaines dans la partie aléatoire. La population est alors considérée comme un échantillon tiré à partir d'une superpopulation infinie régie par le modèle. Cette approche n'est cependant pas très utilisée par les statisticiens des instituts officiels car d'une part, l'utilisation de modèle de superpopulation peut déroger au principe d'impartialité inscrit dans le code de déontologie des statisticiens (voir, par exemple, Tillé, 2001, p. 15), et d'autre part, les méthodes basées sur un modèle fournissent généralement des estimations ponctuelles et non nécessairement des pondérations susceptibles d'être appliquées à d'autres variables d'intérêt. Bien évidemment, l'idée de recourir à un modèle pour les estimations au niveau des petits domaines n'est pas erronée. Il est clair qu'au niveau d'un petit domaine regroupant quelques unités seulement, voire aucune, il peut être nécessaire de supposer l'existence d'un modèle régissant la population afin de pouvoir y fournir des estimations.

Cette thèse se concentre sur des méthodes d'estimation pour petits domaines dont l'approche se range dans la lignée traditionnelle de celle basée sur le plan de sondage. L'inférence statistique est donc uniquement menée en fonction du plan de sondage. L'un des objectifs de cette thèse est de fournir des pondérations à chacune des unités statistiques. La théorie des sondages est, en effet, définie comme un processus d'extrapolation de l'échantillon à la population, et ce, au moyen de l'utilisation des pondérations. Les pondérations constituent donc, pour les statisticiens des instituts officiels, un élément clef pour les estimations de l'ensemble de la population. D'ailleurs, dans certaines littératures hispanophones, le terme *poids* n'est parfois pas traduit par *peso* mais par le terme *factor de expansión*. Cette thèse a également pour but de fournir des méthodes qui respectent la propriété de cohérence selon laquelle la somme des estimations au niveau des domaines est égale à l'estimation globale au niveau de la population. Il serait, en effet, naturellement intuitif de penser que

*(Auvergne, Rhône-Alpes) et la Méditerranée (Languedoc-Roussillon, Provence-Alpes-Côte d'Azur, Corse). Au niveau européen, les* ZEAT *correspondent au niveau 1 de la nomenclature des unités territoriales statistiques."*

plus la somme des estimations au niveau des domaines sera proche de l'estimation globale au niveau de la population, plus la méthode utilisée pour ces estimations sera correcte.

Ce manuscrit de thèse est organisé en cinq chapitres. Les quatre premiers chapitres peuvent être lus indépendamment les uns des autres. Le Chapitre 1 s'ouvre sur quelques éléments de rappel de la théorie statistique des sondages ainsi que sur une introduction à l'estimation sur petits domaines. Il présente les notations et notions importantes en sondage. Le principe de l'inférence statistique y est également rappelé. Ce chapitre donne ensuite un bref panorama des méthodes classiques utilisées pour les petits domaines.

On peut être amené à penser que la variance d'un estimateur par calage diminue lorsque le nombre de variables auxiliaires augmente. Cependant, il peut en être autrement notamment au niveau des domaines. Pour un calage au niveau d'un domaine donné, les poids des unités statistiques peuvent être très dispersés. Cette dispersion s'accroît encore lorsque les variables auxiliaires utilisées sont catégorielles. Un calage sur des variables catégorielles revient en effet à faire un calage sur un jeu de vecteurs composés de zéros et de uns. Le nombre de contraintes de calage ne correspond donc pas au nombre de variables mais au nombre total des modalités des variables. La variance d'un estimateur obtenu par calage est calculée en prenant en compte les résidus de la régression pondérée (avec les poids de calage) de la variable d'intérêt sur les variables auxiliaires. La variance peut donc facilement exploser dès lors que les poids utilisés sont trop dispersés. Le but du Chapitre 2 est de proposer une méthode d'estimation qui consiste à optimiser le redressement d'un sous-échantillon inclus dans un domaine, en utilisant uniquement l'information disponible au niveau du domaine donné. La méthode d'estimation abordée est directe car elle ne fait intervenir que les unités statistiques appartenant au domaine. Ce chapitre est une version reprise d'un article soumis pour publication dans une revue scientifique internationale.

Le Chapitre 3 est une version reprise de Randrianasolo & Tillé (2013). Le chapitre présente la méthode de la scission des poids de sondage. L'idée de la méthode est de créer un poids dépendant à la fois de l'unité et du domaine. Chaque unité peut ainsi contribuer à tous les domaines. Pour une unité donnée, la somme de ses poids relatifs à chaque domaine doit être égale au poids global. Ce poids global peut être l'inverse de la probabilité d'inclusion dans l'échantillon ou un poids résultant d'un calage global au niveau de la population. De plus, on impose que le système de poids pour un domaine particulier satisfasse des propriétés de calage sur les variables auxiliaires connues pour ce domaine. Cette mé-

thode permet alors de construire des estimateurs de type composite qui sont des sommes pondérées d'un estimateur direct et d'un estimateur de type synthétique. Des estimations peuvent être facilement calculées pour n'importe quelle variable d'intérêt, une fois les nouveaux poids calculés.

Le Chapitre 4 propose une réécriture, sous la forme d'estimateurs homogènes linéaires, du célèbre estimateur BLUP (*best linear unbiased predictor* ou meilleur prédicteur linéaire sans biais) de Henderson (1975) ainsi que de sa forme dérivée, l'estimateur pseudo-BLUP proposé par You & Rao (2002), qui utilise les poids de sondage et qui satisfait automatiquement la propriété de cohérence selon laquelle la somme des estimations sur toute partition en domaines est égale à l'estimation globale au niveau de la population. Les nouveaux poids ainsi obtenus à partir de la réécriture des estimateurs BLUP et pseudo-BLUP dépendent néanmoins de constantes de réglage qui dépendent à leur tour de la variable d'intérêt considérée. Le but du Chapitre 4 est alors de proposer des pondérations unifiées pour chacune des unités statistiques pour un jeu de variables d'intérêt données suffisamment corrélées entre elles.

Les méthodes développées dans cette thèse sont ensuite appliquées à l'estimation d'indicateurs de la mobilité locale à partir de l'Enquête Nationale sur les Transports et les Déplacements (ENTD) 2007-2008 dans le Chapitre 5. L'échantillon de l'ENTD 2007-2008 a été tiré dans l'échantillon-maître (EM) de 1999, une réserve d'unités statistiques de l'INSEE, destinée à alimenter les grandes enquêtes nationales auprès des ménages en France. La phase d'échantillonnage de cet EM de 1999 a uniquement pris en compte des niveaux géographiques suffisamment agrégés comme des regroupements de régions, les ZEAT, etc. et non nécessairement des niveaux géographiques plus fins comme les régions. Les domaines considérés dans le Chapitre 5 sont donc les régions. Les différents résultats sur la mobilité locale sont compilés dans l'Annexe A.

# 1 Généralités sur la théorie des sondages et sur l'estimation sur petits domaines

**Résumé** Ce chapitre propose quelques éléments de rappel sur les notions de base de la théorie des sondages et de l'estimation sur petits domaines. La Section 1.1 introduit les notions importantes de la théorie des sondages, dont la notation est largement inspirée de Tillé (2001). La section commence par quelques rappels sur des définitions et notation en théorie des sondages. Une discussion sur les deux différentes approches en sondage est ensuite donnée. La notion d'information auxiliaire est par la suite introduite, pour ainsi aboutir à la présentation de l'estimation par calage. La Section 1.2 est consacrée à une introduction de l'estimation sur petits domaines. Le terme de *petit domaine* y est tout d'abord défini. Plusieurs méthodes classiques pour petits domaines sont ensuite présentées.

**Mots clés** théorie des sondages, petits domaines

## 1.1 Théorie statistique des sondages en populations finies

### 1.1.1 Définitions et notation

Tillé (2001) définit la théorie des sondages comme *"un ensemble d'outils statistiques permettant l'étude d'une population au moyen de l'examen d'une partie de celle-ci"*.

#### Population, caractère et fonction d'intérêt

Considérons $U = \{1, \ldots, k, \ldots, N\}$ une population finie de $N$ unités statistiques. Pour chaque unité de la population, on suppose qu'il existe une information permettant de la repérer. La liste de ces informations pour toutes les unités de la population est ce que l'on appelle *base de sondage*.

L'objet du sondage se porte sur un caractère $y$, appelé *variable d'intérêt*. Ce caractère est supposé observable pour chaque unité $k \in U$ de la population. Malgré son appelation en tant que *variable d'intérêt*, le caractère $y$ n'est pas une variable aléatoire. La valeur prise par la variable d'intérêt $y$ sur l'unité $k$ est notée $y_k$. Cette valeur est fixe et ne sous-entend aucune notion de hasard. Le caractère $y$ peut aussi être vu comme un vecteur de caractères. On considérera dans ce chapitre que $y$ est un caractère. Toutes les définitions qui suivent peuvent donc être étendues et appliquées à des vecteurs de caractères.

Contrairement aux recensements, l'objectif d'un sondage n'est pas de connaître $y$ pour chaque unité d'observation mais plutôt d'estimer une fonction d'intérêt de $y$, notée $\vartheta$, telle que

$$\vartheta = \vartheta(y_k, k \in U).$$

L'on peut, par exemple, s'intéresser à estimer le total des valeurs prises par le caractère $y$ sur toute la population

$$t_y = \sum_{k \in U} y_k,$$

ou encore la moyenne de ces valeurs

$$\overline{y} = \frac{1}{N} \sum_{k \in U} y_k.$$

**Échantillon et plan de sondage**

Un échantillon $s$ est un sous-ensemble de $U$. Soit $\mathcal{Q}$ l'ensemble de tous les échantillons possibles. $\mathcal{Q}$ représente donc l'ensemble des parties non vides de $U$ tel que $\#\mathcal{Q} = 2^N - 1$. Un sous-ensemble qui contient plusieurs fois la même unité d'observation n'est donc pas un échantillon.

**Définition 1.1** *Un plan de sondage $p(.)$ est une loi de probabilité sur l'ensemble $\mathcal{Q}$ de tous les échantillons possibles dans $U$, telle que,*

$$p(s) \geq 0, \text{ et } \sum_{s \in \mathcal{Q}} p(s) = 1.$$

Pour un plan de sondage $p(s)$ donné, un échantillon $s$ est une réalisation d'un échantillon aléatoire $S$, i.e. $\Pr(S = s) = p(s)$ pour tout $s \in \mathcal{Q}$. La taille de l'échantillon $s$ est notée $n(s)$ et est donc de manière générale aléatoire. Lorsque $\mathrm{Var}[n(s)] = 0$, l'échantillon est dit *de taille fixe*. À partir du plan de sondage, la probabilité d'inclusion d'ordre 1 de l'unité $k$, i.e. la probabilité que l'unité $k$ soit présente dans l'échantillon, est donnée par

$$\pi_k = \mathbb{E}(\mathbb{1}_k) = \Pr(k \in S) = \sum_{s \ni k} p(s), \text{ pour tout } k \in U,$$

où $\mathbb{1}_k$ est l'indicatrice d'appartenance de l'unité $k$ à l'échantillon. La probabilité d'inclusion d'ordre 2 des unités $k$ et $l$, i.e. la probabilité que les unités $k$ et $l$ soient toutes deux présentes dans l'échantillon, est donnée par

$$\pi_{kl} = \mathbb{E}(\mathbb{1}_k \mathbb{1}_l) = \Pr(k \in S, l \in S) = \sum_{s \ni k,l} p(s), \text{ pour tout } k, l \in U, k \neq l.$$

Ainsi pour un plan de sondage $p(s)$ donné, on a

$$\mathrm{Var}(\mathbb{1}_k) = \mathbb{E}(\mathbb{1}_k^2) - \mathbb{E}(\mathbb{1}_k)^2 = \pi_k(1 - \pi_k), \text{ pour tout } k \in U,$$

et pour tout $k, l \in U, k \neq l$,

$$\mathrm{Cov}(\mathbb{1}_k, \mathbb{1}_l) = \mathbb{E}(\mathbb{1}_k \mathbb{1}_l) - \mathbb{E}(\mathbb{1}_k)\mathbb{E}(\mathbb{1}_l) = \pi_{kl} - \pi_k \pi_l.$$

De plus, on peut noter

$$\Delta_{kl} = \begin{cases} \mathrm{Cov}(\mathbb{1}_k, \mathbb{1}_l) & \text{si } k \neq l, \\ \mathrm{Var}(\mathbb{1}_k) & \text{sinon.} \end{cases}$$

Il existe de nombreux plans de sondage, nous en présenterons seulement cinq qui seront utilisés dans la suite de ce document : le plan de sondage aléatoire simple sans remise, le plan de sondage stratifié, le plan de Poisson, le plan de sondage à deux phases et le plan de sondage équilibré.

**Définition 1.2** *Un plan de sondage aléatoire simple sans remise de taille n est un plan de taille fixe n tel que*

$$p(s) = \begin{cases} \binom{N}{n}^{-1} & \textit{si } s \textit{ est de taille égale à } n, \\ 0 & \textit{sinon.} \end{cases}$$

La probabilité d'inclusion d'ordre 1 du plan de sondage aléatoire simple sans remise est

$$\pi_k = \sum_{s \ni k} p(s) = \sum_{s \ni k} \binom{N}{n}^{-1} = \frac{n}{N}, \text{ pour tout } k \in U,$$

et la probabilité d'ordre 2 est

$$\pi_{kl} = \sum_{s \ni k,l} p(s) = \sum_{s \ni k,l} \binom{N}{n}^{-1} = \frac{n(n-1)}{N(N-1)}, \text{ pour tout } k \neq l \in U.$$

Enfin, il vient, pour tout $k, l \in U$,

$$\Delta_{kl} = \begin{cases} -\dfrac{n(N-n)}{N^2(N-1)} & \text{si } k \neq l, \\ \dfrac{n(N-n)}{N^2} & \text{sinon.} \end{cases}$$

**Définition 1.3** *Supposons que la population $U$ soit partitionnée en $H$ sous-ensembles, $U_h, h = 1, \ldots, H$, appelés strates, de tailles respectives $N_h, h = 1, \ldots, H$, supposées connues. Un sondage est dit stratifié si, dans chaque strate $U_h$, un échantillon aléatoire $S_h$ est sélectionné suivant un plan de sondage aléatoire simple sans remise de taille $n_h$, et si la sélection d'un échantillon dans une strate est indépendante de la sélection d'un échantillon dans toutes les autres strates. En notant $s_h$ une valeur possible de l'échantillon aléatoire $S_h$ prévelé dans la strate $U_h$ suivant le plan $p_h(.)$ avec $p_h(s_h) = Pr(S_h = s_h)$, le plan de sondage stratifié est donné par*

$$p(s) = \prod_{h=1}^{H} p_h(s_h), s = \bigcup_{h=1}^{H} s_h.$$

Il en découle que la probabilité d'inclusion d'ordre 1 pour un plan de sondage stratifié est

$$\pi_k = \frac{n_h}{N_h}, \qquad \text{pour } k \in U_h.$$

La probabilité d'inclusion d'ordre 2 est

$$\pi_{kl} = \frac{n_h(n_h - 1)}{N_h(N_h - 1)}, \qquad \text{si } k, l \in U_h,$$

et

$$\pi_{kl} = \frac{n_h n_i}{N_h N_i}, \qquad \text{si } k \in U_h \text{ et } l \in U_i.$$

Enfin, il en découle

$$\Delta_{kl} = \begin{cases} \dfrac{n_h}{N_h}\dfrac{N_h - n_h}{N_h} & \text{si } k = l, k \in U_h, \\ -\dfrac{n_h(N_h - n_h)}{N_h^2(N_h - 1)} & \text{si } k, l \in U_h, K \neq l, \\ 0 & \text{sinon.} \end{cases}$$

**Définition 1.4** *Un plan de Poisson est un plan de sondage donné par*

$$p(s) = \prod_{k \in s} \pi_k \prod_{k \in U \setminus s} (1 - \pi_k), \text{ pour tout } s \subset U,$$

*où $\pi_k$ est la probabilité d'inclusion d'ordre 1, pour $k \in U$.*

L'échantillon obtenu par un plan de Poisson est composé de toutes les unités de la population $U$ satisfaisant à l'inégalité $u_k < \pi_k$ où $u_k$ est la réalisation indépendante d'une variable aléatoire de loi uniforme sur l'intervalle $[0, 1]$, pour $k \in U$. Les unités étant sélectionnées indépendamment les unes des autres, on a $\pi_{kl} = \pi_k \pi_l$, pour tout $k \neq l$. La taille de l'échantillon est aléatoire avec $\mathbb{E}[n(s)] = \sum_{k \in U} \pi_k$ et $\text{Var}[n(s)] = \sum_{k \in U} \pi_k(1 - \pi_k)$.

Si on veut des probabilités d'inclusion d'ordre 1, $\pi_k, k \in U$, proportionnelles à certains nombres strictement positifs $x_k, k \in U$, connus dans

la population et une taille espérée de l'échantillon égale à $n$, on peut prendre

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

Il est clair que les probabilités d'inclusion $\pi_k, k \in U$ peuvent prendre des valeurs supérieures à un. Afin d'y remédier, les unités statistiques ayant des probabilités d'inclusion plus grandes que un sont automatiquement sélectionnées dans l'échantillon. Les probabilités d'inclusion des unités non sélectionnées d'office sont recalculées de la même manière tout en reconsidérant que la nouvelle taille de l'échantillon est diminuée du nombre des unités ayant les probabilités d'inclusion supérieures à un. Cette procédure est répétée jusqu'à ce que toutes les probabilités d'inclusion soient égales à un, ou soient strictement proportionnelles aux nombres $x_k, k \in U$.

**Définition 1.5** *Un plan de sondage à deux phases consiste à sélectionner un échantillon issu d'un échantillon de la population. La première phase consiste à tirer un échantillon aléatoire $S_a$ dans la population U au moyen d'un plan de sondage quelconque $p(s_a) = Pr(S_a = s_a)$, de taille non nécessairement fixe. La deuxième consiste à tirer un échantillon aléatoire $S_b$ à partir de l'échantillon aléatoire $S_a$ de la première phase, selon un autre plan de sondage $p(s_b|S_a) = Pr(S_b = s_b|S_a)$.*

Les probabilités d'inclusion d'ordres 1 et 2 de la première phase sont notées :

$$\pi_{ak} = \Pr(k \in S_a) = \sum_{s_a \ni k} p(s_a), \text{ pour tout } k \in U,$$

et

$$\pi_{akl} = \Pr(k \in S_a, l \in S_a) = \sum_{s_a \ni k,l} p(s_a), \text{ pour tout } k, l \in U, k \neq l,$$

avec $\pi_{akk} = \pi_{ak}$.
Pour la deuxième phase, les probabilités d'inclusion d'ordres 1 et 2 sont données par :

$$\pi_{bk} = \Pr(k \in S_b|S_a) = \sum_{s_b \ni k} p(s_b|S_a), \text{ pour tout } k \in S_a,$$

et

$$\pi_{bkl} = \Pr(k \in S_b|S_a, l \in S_b|S_a) = \sum_{s_b \ni k,l} p(s_b|S_a), \text{ pour tout } k, l \in S_a, k \neq l,$$

avec $\pi_{bkk} = \pi_{bk}$.

**Définition 1.6** *Considérons un vecteur $\mathbf{z}_k = (z_{k1}, \ldots, z_{kp}, \ldots, z_{kP})^\top$ de valeurs prises par P caractères auxiliaires pour une unité k. Le total*

$$\mathbf{t}_z = \sum_{k\in U} \mathbf{z}_k$$

*est supposé connu. Un plan de sondage $p(s)$ est équilibré sur les caractères auxiliaires $z_1, \ldots, z_P$ si et seulement si les équations d'équilibrage*

$$\sum_{k \in s} \frac{\mathbf{z}_k}{\pi_k} = \sum_{k \in U} \mathbf{z}_k = \mathbf{t}_z$$

*sont vérifiées pour tout $s \subset U$ tel que $p(s) > 0$.*

La méthode du Cube proposée par Deville & Tillé (2004) permet de sélectionner un échantillon suivant un plan de sondage équilibré.

### 1.1.2 Inférence statistique en théorie des sondages

Dans la théorie des sondages, deux grandes approches pour réaliser des inférences s'affrontent : l'approche basée sur le plan de sondage ou approche par "randomisation" d'une part, et l'approche basée sur un modèle ou approche par prédiction d'autre part. Dans la première approche, l'inférence est basée sur la distribution de probabilité utilisée pour sélectionner l'échantillon. Dans la deuxième approche, la population finie $U$ est traitée comme étant un simple échantillon tiré à partir d'un modèle dit de superpopulation. La distribution du modèle forme ainsi la base de l'inférence. Les principales divergences entre ces deux différentes écoles de pensée ont été largement discutées par plusieurs auteurs (voir, par exemple, Brewer, 1994, 1999; Brewer et al., 1988; Cassel et al., 1977; Hansen et al., 1983; Little, 2004; Nedyalkova & Tillé, 2008; Royall, 1988; Särndal, 1978; Smith, 1994). Une tentative de synthèse a toutefois été developpée (Särndal et al., 1992), ayant abouti à une approche dite "assistée par un modèle" qui permet de fournir des inférences valides même lorsque le modèle est faux. En effet, Särndal et al. (1992, p. 227) précise que le rôle du modèle est seulement de *"décrire le nuage de points de la population finie"*. Selon les auteurs, *"la population finie paraît comme pouvant être générée conformément à un modèle. Cependant l'hypothèse selon laquelle la population finie est réellement générée à partir d'un modèle n'est jamais considérée"*. Cette nouvelle approche hydride est vue comme une approche basée sur le plan de sondage car l'approche n'est pas dépendante du modèle mais assistée par celui-ci.

#### L'approche basée sur le plan

L'approche basée sur le plan peut être vue comme l'approche classique (ou traditionnelle) en théorie des sondages. Cochran (1977) et Kish (1965) en donnent par exemple une description complète dans leurs livres. Dans cette approche, les valeurs $y_k, k \in U$ de la variable d'intérêt dans la population sont supposées fixes. L'objectif est de fournir un estimateur, noté $\hat{\vartheta}$, de $\vartheta$, qui serait sans biais sous le plan. Plus précisément, l'estimateur $\hat{\vartheta}$ devrait être noté $\hat{\vartheta}(s)$ car il ne dépend que du plan de sondage.

**Définition 1.7** *Un estimateur $\hat{\vartheta}$ de $\vartheta$ est sans biais sous le plan si l'espérance de $\hat{\vartheta}$ par rapport au plan de sondage est égale à $\vartheta$, i.e.*

$$\mathbb{E}_p(\hat{\vartheta}) = \sum_{s \in \mathcal{Q}} p(s)\hat{\vartheta}(s) = \vartheta.$$

La variance est un indicateur important de la précision d'un estimateur donné et permet de construire des intervalles de confiance.

**Définition 1.8** *La variance sous le plan d'un estimateur $\hat{\vartheta}$ de $\vartheta$ est définie par*

$$\mathrm{Var}_p(\hat{\vartheta}) = \sum_{s \in \mathcal{Q}} p(s)[\hat{\vartheta}(s) - \mathbb{E}_p(\hat{\vartheta})]^2 = \mathbb{E}_p[\hat{\vartheta} - \mathbb{E}_p(\hat{\vartheta})]^2.$$

L'estimateur le plus utilisé dans cette approche est celui de Horvitz & Thompson (1952), aussi connu sous le nom de "$\pi$-estimateur". Les observations sont pondérées par l'inverse des probabilités d'inclusion d'ordre 1. L'estimateur de Horvitz & Thompson (1952) du total $t_y = \sum_{k \in U} y_k$ est alors donné par

$$\hat{t}_{y,\pi} = \sum_{k \in S} \frac{y_k}{\pi_k},$$

en supposant $\pi_k > 0$, pour tout $k \in U$. Cet estimateur du total obtenu a la particularité d'être sans biais sous le plan de sondage.

Toujours en supposant $\pi_k > 0$, pour tout $k \in U$, la variance de $\hat{t}_{y,\pi}$ est donnée par

$$\begin{aligned} \mathrm{Var}_p(\hat{t}_{y,\pi}) &= \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{y_k y_l}{\pi_k \pi_l}, \\ &= \sum_{k \in U} \frac{\pi_k(1-\pi_k)}{\pi_k^2} y_k^2 + \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} y_k y_l. \end{aligned}$$

De même, Horvitz & Thompson (1952) ont proposé un estimateur sans biais de cette variance dans le cas où $\pi_{kl} > 0$, pour tout $k, l \in U$,

$$\widehat{\mathrm{Var}}_p(\hat{t}_{y,\pi}) = \sum_{k \in S} \frac{1-\pi_k}{\pi_k^2} y_k^2 + \sum_{k \in S} \sum_{\substack{l \in S \\ l \neq k}} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l} y_k y_l.$$

Cette variance a cependant l'inconvénient de parfois prendre des valeurs négatives. En considérant un plan de sondage de taille fixe et en supposant $\pi_k > 0$ pour tout $k \in U$, Yates & Grundy (1953) ainsi que Sen (1953) ont proposé indépendamment d'autres expressions différentes de la variance et de l'estimateur de la variance de $\hat{t}_{y,\pi}$. Ils ont en effet montré qu'il était possible de formuler la variance de l'estimateur de Horvitz & Thompson (1952) par :

$$\mathrm{Var}_p(\hat{t}_{y,\pi}) = -\frac{1}{2} \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \Delta_{kl}.$$

L'estimateur Sen-Yates-Grundy de la variance de $\hat{t}_{y,\pi}$ est donné par :

$$\widehat{\mathrm{Var}}_p^{SYG}(\hat{t}_{y,\pi}) = -\frac{1}{2}\sum_{k\in S}\sum_{\substack{l\in S\\ l\neq k}}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2\frac{\Delta_{kl}}{\pi_{kl}}.$$

Lorsque le plan est de taille fixe, l'estimateur Sen-Yates-Grundy de la variance est sans biais. Une condition suffisante pour qu'il soit positif est d'avoir l'inégalité $\pi_k\pi_l - \pi_{kl} \geq 0$, pour tout $k, l \in U$, $k \neq l$.

Lorsque la taille de l'échantillon considéré est assez grande, on suppose que l'estimateur de Horvitz & Thompson (1952) suit approximativement une loi normale. Un intervalle de confiance au niveau de confiance $1-\alpha$ pour le total $t_y$ est donc :

$$IC_{1-\alpha} = \left[\hat{t}_{y,\pi} - z_{1-\alpha/2}\sqrt{\mathrm{Var}_p(\hat{t}_{y,\pi})}, \hat{t}_{y,\pi} + z_{1-\alpha/2}\sqrt{\mathrm{Var}_p(\hat{t}_{y,\pi})}\right],$$

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

**L'approche basée sur un modèle**

Dans l'approche basée sur le modèle, les valeurs prises par la variable d'intérêt sur les unités d'observation de la population sont supposées être des réalisations de variables aléatoires. La population est donc vue comme étant une réalisation fournie par un modèle, celui d'une superpopulation. Comme l'évoque Tillé (1992, 2001), cette idée est née de Brewer (1963) et a été par la suite formalisée et développée par Royall (1970, 1971, 1976). Le livre de Valliant et al. (2000) est devenue une référence en ce qui concerne la théorie de l'approche basée sur un modèle.

Prenons l'exemple d'un modèle linéaire $\xi$ de superpopulation tel que :

$$\xi : y_k = x_k\beta + \varepsilon_k,$$

avec $\mathbb{E}_\xi(\varepsilon_k) = 0, \mathrm{Var}_\xi(\varepsilon_k) = \kappa_k^2\sigma^2$ et $\mathrm{Cov}_\xi(\varepsilon_k, \varepsilon_l) = 0$, pour $k, l \in U, k \neq l$. $\mathbb{E}_\xi(.), \mathrm{Var}_\xi(.)$ et $\mathrm{Cov}_\xi(.,.)$ représentent l'espérance, la variance et la covariance par rapport au modèle $\xi$. Les $x_k, k \in U$ ne sont pas aléatoires et sont supposés connus. De même, les quantités $\kappa_k^2, k \in U$ modélisant une éventuelle hétéroscédasticité sont supposées connues.

Soit $S$ un échantillon de la population $U$ et soit $U\backslash S$ l'ensemble des unités de la population $U$ qui n'appartiennent pas à l'échantillon. Le total de la variable d'intérêt $y$ dans la population peut alors s'écrire comme

$$t_y = \sum_{k\in S} y_k + \sum_{k\in U\backslash S} y_k.$$

Si le paramètre $\beta$ est connu, nous pouvons estimer le total par

$$\hat{t}_{y,\xi} = \sum_{k \in S} y_k + \beta \sum_{k \in U \setminus S} x_k.$$

Dans le cas contraire, l'estimateur du total est

$$\hat{t}_{y,\xi} = \sum_{k \in S} y_k + \hat{\beta} \sum_{k \in U \setminus S} x_k,$$

où $\hat{\beta}$ est obtenu par la méthode du maximum de vraisemblance.

Tout comme précédemment, l'objectif est ici de fournir un estimateur, noté $\hat{\vartheta}$, de $\vartheta$, qui serait sans biais, mais cette fois-ci, sous le modèle $\xi$.

**Définition 1.9** *Un estimateur $\hat{\vartheta}$ de $\vartheta$ est sans biais sous le modèle si l'espérance de $\hat{\vartheta}$ par rapport au modèle $\xi$ est égale à $\vartheta$, i.e.*

$$\mathbb{E}_\xi(\hat{\vartheta}) = \vartheta.$$

**Définition 1.10** *La variance sous le modèle d'un estimateur $\hat{\vartheta}$ de $\vartheta$ est définie par*

$$\mathrm{Var}_\xi(\hat{\vartheta}) = \mathbb{E}_\xi[\hat{\vartheta} - \mathbb{E}_\xi(\hat{\vartheta})]^2.$$

Dans l'approche basée sur le modèle, il est important de bien vérifier la validité du modèle utilisée. Dans ce cas, l'utilisation du modèle permet d'obtenir des estimations optimales sur les critères du biais ainsi que de l'erreur quadratique moyenne. Dans le cas contraire, les estimations obtenues seraient peu fiables.

#### Quelle approche choisir ?

Le Tableau 1.1 est un tableau comparatif des deux différentes approches en théorie des sondages.

L'approche utlisée par la suite dans ce document est l'approche basée sur le plan de sondage, les travaux qui suivent étant axé sur l'utilisation et la manipulation de poids de sondage. Ainsi, les espérances et variances considérées dans la suite de ce document seront donc des espérances et des variances par rapport au plan de sondage. Sauf mention contraire, il convient donc de considérer les notations $\mathbb{E}(.), \mathrm{Var}(.)$ et $\mathrm{Cov}(.)$ comme étant équivalentes à $\mathbb{E}_p(.), \mathrm{Var}_p(.)$ et $\mathrm{Cov}_p(.)$.

### 1.1.3 Information auxiliaire

Une variable est dite auxiliaire lorsqu'elle est disponible avant même la réalisation d'une enquête. L'information qu'elle véhicule est jugée complète. Elle permet donc d'améliorer la précision des estimations (Ardilly, 2006a; Ardilly & Tillé, 2003, 2005; Särndal et al., 1992; Tillé, 1992,

TABLE 1.1 – *Inférence sous un modèle vs inférence sous le plan. Tableau tiré de Molina & Rao (2012, p. 4).*

| Éléments | Approches | |
|---|---|---|
| | sur un modèle | sur le plan |
| Population | $\mathbf{Y} \sim \xi$ | $U = \{1, \ldots, k, \ldots, N\}$<br>$y = \{y_1, \ldots, y_n\}$ |
| Échantillon | $y = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_k, \ldots, \mathbf{Y}_n\}$<br>$\mathbf{Y}_k$ i.i.d comme $\mathbf{Y}$ | $s = \{1, \ldots, k, \ldots, n\} \in \mathcal{Q}$<br>$y = \{y_1, \ldots, y_k, \ldots, y_n\}$ |
| Loi de probabilité | $p_\xi(y)$ | $p(s) = \Pr(S = s)$ |
| Paramètre | $\vartheta = \mathbb{E}_\xi(\mathbf{Y})$ | $\vartheta = \vartheta(y_k, k \in U)$ |
| Estimateur | $\hat{\vartheta}(y)$ | $\hat{\vartheta}(s)$ |

2001). Une variable auxiliaire peut être quantitative ou catégorielle. L'information auxiliaire peut regrouper un ensemble de variables. On parlera alors d'un ensemble ou d'un jeu de variables auxiliaires. Les recensements, mais aussi les registres de population par exemple, fournissent des informations disponibles pour toute la population considérée. Ils peuvent donc être source d'information auxiliaire. Considérons un vecteur $\mathbf{x}_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kJ})^\top$ de valeurs prises par $J$ variables auxiliaires pour une unité $k$. Le total

$$\mathbf{t}_x = \sum_{k \in U} \mathbf{x}_k$$

est supposé connu. La connaissance des nombres $\mathbf{x}_k, k \in U$ dans la population peut être utilisée pour créer un plan de sondage, comme par exemple, créer des probabilités d'inclusion d'ordre 1 qui seraient proportionnelles aux $\mathbf{x}_k, k \in U$ (voir, par exemple, la Définition 1.4). Enfin, les variables auxiliaires peuvent aussi être directement utilisées dans les formules des estimateurs.

### 1.1.4 Estimation par calage

L'idée générale de la méthode d'estimation par calage a été développée et formalisée par Deville & Särndal (1992), bien que de nombreux anciens travaux utilisaient déjà les méthodes d'ajustements de tableaux à des marges connues (Deming & Stephan, 1940; Lemel, 1976; Madre, 1979, 1980; Stephan, 1942). Disposant d'un vecteur de variables auxiliaires $\mathbf{x}_k, k \in U$, étant correlées avec la variable d'intérêt $y$, et avec un vecteur de totaux $\mathbf{t}_x = \sum_{k \in U} \mathbf{x}_k$ supposé connu, le but est d'estimer le total $t_y$ en utilisant l'information donnée par $\mathbf{t}_x$.

L'estimateur par calage du total $t_y = \sum_{k \in U} y_k$ s'écrit

$$\hat{t}_{y,w} = \sum_{k \in S} w_k y_k,$$

où les poids $w_k(S), k \in S$ affectés aux unités vérifient les équations de calage

$$\sum_{k \in S} w_k \mathbf{x}_k^\top = \mathbf{t}_x.$$

Puisqu'il existe une infinité de poids $w_k$ qui satisfont ces contraintes, le but est de chercher les poids les *plus proches* des poids $d_k = 1/\pi_k$, l'estimateur de Horvitz & Thompson (1952) étant sans biais.

On choisit alors une fonction de pseudo-distance $G_k(w_k, d_k)$ (voir Tableau 1.2). La fonction $G_k(w_k, d_k)$ est supposée positive, dérivable par rapport à $w_k$, strictement convexe, telle que $G_k(d_k, d_k) = 0$. Le problème consiste ensuite à déterminer les poids $w_k, k \in S$ qui minimisent la quantité

$$\sum_{k \in S} \frac{G_k(w_k, d_k)}{q_k}$$

sous les contraintes de calage

$$\sum_{k \in S} w_k \mathbf{x}_k^\top = \mathbf{t}_x,$$

où les inverses des poids $q_k, k \in S$ sont des coefficients de pondération qui permettent de déterminer l'importance de chaque unité dans le calcul de distance.

TABLE 1.2 – *Quelques pseudo-distances pour le calage. Tableau tiré de Deville & Särndal (1992, Table 1).*

| $G_k(w_k, d_k)$ | $g_k(w_k, d_k)$ | $F_k(u)$ | Type |
|---|---|---|---|
| $\frac{(w_k - d_k)^2}{2d_k}$ | $\frac{w_k}{d_k} - 1$ | $1 + q_k u$ | Khi-deux |
| $w_k \log \frac{w_k}{d_k} + d_k - w_k$ | $\log \frac{w_k}{d_k}$ | $\exp(q_k u)$ | Entropie |
| $2(\sqrt{w_k} - \sqrt{d_k})^2$ | $2\left(1 - \sqrt{\frac{d_k}{w_k}}\right)$ | $(1 - q_k u/2)^{-2}$ | Distance de Hellinger |
| $d_k \log \frac{d_k}{w_k} + w_k - d_k$ | $1 - \frac{d_k}{w_k}$ | $(1 - q_k u)^{-1}$ | Entropie inverse |
| $\frac{(w_k - d_k)^2}{2w_k}$ | $\frac{1}{2}\left(1 - \frac{d_k^2}{w_k^2}\right)$ | $(1 - 2q_k u)^{-1/2}$ | Khi-deux inverse |

Les poids $w_k, k \in S$ sont alors définis par

$$w_k = d_k F_k(q_k \mathbf{x}_k \boldsymbol{\lambda}),$$

$F_k(.)$ représentant l'inverse de la fonction $g_k(w_k, d_k)$ qui est la dérivée de $G_k(w_k, d_k)$ par rapport à $w_k$, $\boldsymbol{\lambda}$ étant le multiplicateur de Lagrange découlant des contraintes

$$\mathbf{t}_x = \sum_{k \in S} d_k \mathbf{x}_k^\top F(q_k \mathbf{x}_k \boldsymbol{\lambda}).$$

En prenant la pseudo-distance de type *Khi-deux*, l'estimateur par calage coïncide avec l'estimateur par la régression généralisée

$$\hat{t}_{y,reg} = \hat{t}_{y,\pi} + (\mathbf{t}_x - \mathbf{t}_{x,\pi})\widehat{\mathbf{B}},$$

avec

$$\widehat{\mathbf{B}} = \left( \sum_{k \in S} \frac{\mathbf{x}_k \mathbf{x}_k^\top q_k}{\pi_k} \right)^{-1} \sum_{k \in S} \frac{\mathbf{x}_k y_k q_k}{\pi_k}.$$

## 1.2 Introduction à l'estimation sur petits domaines

En sondage, l'objectif n'est pas toujours d'estimer une fonction d'intérêt au niveau de la population $U$. On peut aussi s'intéresser à estimer la fonction d'intérêt pour une sous-population $U_0 \subset U$ appelée *domaine*. Le terme *domaine* peut donc se référer à une zone géographique donnée (on utilisera en anglais le terme *area*), à un groupe socio-démographique, une catégorie ou un croisement de plusieurs variables catégorielles (on parlera alors de *domain* en anglais, ou certaines fois, par abus de langage, *area*). Ainsi, un *domaine* peut désigner un département, une région, les personnes au chômage, les personnes de plus de 65 ans, les personnes à mobilité réduite, les personnes de formation mathématique habitant la région d'Alsace, l'ensemble des mots contenant la lettre *e* dans le roman *La Disparition* de Perec (1989), etc.

Särndal et al. (1992) définissent un *petit domaine* comme étant un *domaine* qui ne regroupe qu'une infime fraction de la population. Rao (2003) et Ardilly (2006b) considèrent qu'un *domaine* est vu comme *petit*, lorsqu'une estimation classique relative au *domaine*, a une précision insuffisante. Plus intuitivement, un *petit domaine* peut être vu comme une sous-population dont la taille est faible, voire nulle, dans l'échantillon. La forte demande de résultats d'enquêtes nationales, notamment au niveau géographique fin, a soulevé le problème de l'*estimation sur petits domaines* (ou en anglais *small domain/area estimation* ou uniquement *small area estimation*). Les techniques d'*estimation sur petits domaines* se sont beaucoup développées depuis les années 1970, notamment grâce à l'émergence d'une nouvelle école de pensée en théorie des sondages, l'approche basée sur un modèle. Quelques méthodes classiques d'*estimation sur petits domaines* sont présentées dans la suite de ce chapitre, la présentation sera largement inspirée de Chauvet (2011); Lahiri et al. (2011); Molina & Rao (2012); Särndal et al. (1992).

### 1.2.1 Méthodes traditionnelles utilisant l'approche basée sur le plan de sondage

Considérons une partition de la population $U = \{1, \ldots, k, \ldots, N\}$ en $D$ domaines $U_d, d = 1, \ldots, D$, de tailles respectives $N_d, d = , \ldots, D$. Par exemple, les strates définies dans la Définition 1.3 peuvent être vue comme des domaines. L'objectif est maintenant d'estimer une fonction d'intérêt de $y$, non plus au niveau de la population $U$, mais au niveau des domaines

$U_d, d = 1, \ldots, D$, notée $\vartheta_d$, telle que

$$\vartheta_d = \vartheta_d(y_k, k \in U_d), \text{ pour } d = 1, \ldots, D.$$

Par la suite, nous ne considérerons que le total

$$t_y^d = \sum_{k \in U_d} y_k$$

ou la moyenne

$$\overline{y}_d = \frac{1}{N_d} \sum_{k \in U_d} y_k$$

comme fonctions d'intérêt.

Dans la pratique, la taille $N_d$ est souvent inconnue. Estimer la moyenne $\overline{y}_d$ revient donc à estimer le ratio de deux paramètres inconnus. Dans la suite de ce document, les tailles $N_d, d = 1, \ldots, D$ seront toujours supposées connues.

Supposons maintenant qu'un échantillon aléatoire $S$ de taille $n$ est tiré dans la population suivant un plan de sondage donné $p(s)$ avec des probabilités d'inclusion $\pi_k, \pi_{kl}$, pour $k, l \in U$. Notons $n_d$ les tailles des domaines dans l'échantillon $S_d = S \cap U_d$ pour $d = 1, \ldots, D$. Les tailles $n_d, d = 1, \ldots, D$ sont aléatoires et peuvent être très petites, voire nulles.

### Estimation directe

**Définition 1.11** *Une estimation au niveau d'un domaine est directe lorsqu'elle est construite en n'utilisant aucune information extérieure au domaine.*

Au niveau d'un domaine donné, l'estimateur de Horvitz & Thompson (1952) et l'estimateur par calage (dont l'estimateur par la régression en est un cas particulier) sont des estimateurs directs.

**Estimateur de Horvitz & Thompson (1952)** L'estimateur de Horvitz & Thompson (1952) du total $t_y^d = \sum_{k \in U_d} y_k$ est alors donné par

$$\hat{t}_{y,\pi}^d = \sum_{k \in S_d} \frac{y_k}{\pi_k},$$

en supposant $\pi_k > 0$, pour tout $k \in U_d$.

Dans le cas d'un plan de sondage à taille fixe, la variance est donnée par

$$\mathrm{Var}(\hat{t}_{y,\pi}^d) = -\frac{1}{2} \sum_{k \in U_d} \sum_{\substack{l \in U_d \\ l \neq k}} \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \Delta_{kl}.$$

Un estimateur de la variance est donné par :

$$\widehat{\text{Var}}(\hat{t}^d_{y,\pi}) = -\frac{1}{2}\sum_{k\in S_d}\sum_{\substack{l\in S_d\\ l\neq k}}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2\frac{\Delta_{kl}}{\pi_{kl}}.$$

Lorsque les tailles $N_d, d = 1,\ldots,D$ sont connues, un meilleur estimateur du total $t^d_y$, dérivé de celui de Horvitz & Thompson (1952), est donné par

$$\tilde{t}^d_y = N_d\tilde{y}_d = \frac{N_d}{\hat{N}_d}\sum_{k\in S_d}\frac{y_k}{\pi_k},$$

avec $\hat{N}_d = \sum_{k\in S_d}\pi_k^{-1}$. Une approximation de la variance est donnée par

$$\text{AVar}(\tilde{t}^d_y) = \sum_{k\in U_d}\sum_{l\in U_d}\left(\frac{y_k-\overline{y}_d}{\pi_k}\right)\left(\frac{y_l-\overline{y}_d}{\pi_l}\right)\Delta_{kl}.$$

Un estimateur de la variance est donné par

$$\widehat{\text{Var}}(\tilde{t}^d_y) = \left(\frac{N_d}{\hat{N}_d}\right)^2\sum_{k\in S_d}\sum_{l\in S_d}\left(\frac{y_k-\tilde{y}_d}{\pi_k}\right)\left(\frac{y_l-\tilde{y}_d}{\pi_l}\right)\frac{\Delta_{kl}}{\pi_{kl}}.$$

**Estimateur par calage** L'estimateur par calage du total $t^d_y = \sum_{k\in U_d} y_k$ est alors donné par

$$\hat{t}^d_{y,w1} = \sum_{k\in S_d} w_{1k}y_k,$$

où les poids $w_{1k}, k \in S_d$ résultent d'un calage sur les totaux des variables auxiliaires au niveau du domaine $U_d$. En utilisant les techniques des résidus (Deville & Särndal, 1992; Deville et al., 1993; Tillé, 2001), la variance est donnée par

$$\begin{aligned}\text{AVar}(\hat{t}^d_{y,w1}) &\simeq \text{Var}(\hat{t}^d_E)\\ &\simeq \sum_{k\in U_d}\sum_{l\in U_d}\frac{E_k}{\pi_k}\frac{E_l}{\pi_l}\Delta_{kl},\end{aligned}$$

où $E_k = y_k - \mathbf{x}_k^\top\mathbf{B}^d_1$ donne les résidus de la régression de $y$ sur le jeu des variables auxiliaires au niveau de la sous-population $U_d$.

Un estimateur de la variance est alors donné par

$$\widehat{\text{Var}}(\hat{t}^d_{y,w1}) = \sum_{k\in S_d}\sum_{l\in S_d}\frac{\Delta_{kl}}{\pi_{kl}}w_{1k}e_kw_{1l}e_l,$$

où $e_k = y_k - \mathbf{x}_k^\top\hat{\mathbf{B}}^d_{1s}$ donne les résidus de la régression pondérée de $y$ sur le jeu des variables auxiliaires au niveau du sous-échantillon $S_d$.

Une autre variante de l'estimateur par calage du total $t_y^d$ est donné par

$$\hat{t}_{y,w2}^d = \sum_{k \in S_d} w_{2k} y_k,$$

où les poids $w_{2k}, k \in S$ résultent d'un calage sur les totaux des variables auxiliaires au niveau de la population $U$. Une approximation de la variance de l'estimateur est donnée par

$$\begin{aligned} \mathrm{AVar}(\hat{t}_{y,w2}^d) &\simeq \mathrm{Var}(\hat{t}_E^d) \\ &\simeq \sum_{k \in U} \sum_{l \in U} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l} \Delta_{kl}, \end{aligned}$$

où, cette fois-ci, $E_k = y_k \mathbb{1}(k \in U_d) - \mathbf{x}_k^\top \mathbf{B}_2^d$ donne les résidus de la régression de $y\mathbb{1}(U_d)$ sur le jeu des variables auxiliaires au niveau de la population $U$.

Un estimateur de la variance est alors donné par

$$\widehat{\mathrm{Var}}(\hat{t}_{y,w2}^d) = \sum_{k \in S} \sum_{l \in S} \frac{\Delta_{kl}}{\pi_{kl}} w_{2k} e_k w_{2l} e_l,$$

où $e_k = y_k \mathbb{1}(k \in S_d) - \mathbf{x}_k^\top \hat{\mathbf{B}}_{2s}^d$ donne les résidus de la régression pondérée de $y\mathbb{1}(S_d)$ sur le jeu des variables auxiliaires au niveau de l'échantillon $S$.

Les estimateurs directs du total $t_y^d$ ont la particularité d'avoir un faible biais mais une variance de l'ordre de $\mathcal{O}(1/n_d)$. Plus la taille $n_d$ est faible, plus la variance augmente. Il est donc nécessaire d'*emprunter de la force ailleurs*.

### Estimation indirecte

**Définition 1.12** *Une estimation au niveau d'un domaine est indirecte lorsqu'elle est construite en utilisant de l'information provenant également de l'extérieur du domaine.*

**Définition 1.13** *Une estimation au niveau d'un domaine est synthétique lorsqu'un paramètre défini pour la population globale est supposé être le même pour le domaine.*

**Estimateur synthétique** Disposant de l'information auxiliaire, l'estimateur indirect synthétique du total $t_y^d = \sum_{k \in U_d} y_k$ s'écrit alors

$$\hat{t}_{y,syn}^d = \mathbf{t}_x^{d\top} \hat{\mathbf{B}}_s,$$

où $\hat{\mathbf{B}}_s$ est un paramètre estimé au niveau de tout l'échantillon $S$, tel que

$$\hat{\mathbf{B}}_s = \left( \sum_{k \in S} \frac{\mathbf{x}_k \mathbf{x}_k^\top q_k}{\pi_k} \right)^{-1} \sum_{k \in S} \frac{\mathbf{x}_k y_k q_k}{\pi_k}.$$

Le biais de l'estimateur synthétique est faible si les variables auxiliaires expliquent bien la variable d'intérêt, et dans l'hypothèse *implicite* que pour $d = 1, \ldots, D$,

$$\begin{aligned} \mathbf{B}^d &\simeq \mathbf{B}, \\ &\text{i.e.} \\ \left(\sum_{k\in U_d} \mathbf{x}_k\mathbf{x}_k^\top q_k\right)^{-1} \sum_{k\in U_d} \mathbf{x}_k y_k q_k &\simeq \left(\sum_{k\in U} \mathbf{x}_k\mathbf{x}_k^\top q_k\right)^{-1} \sum_{k\in U} \mathbf{x}_k y_k q_k. \end{aligned}$$

Le biais pouvant être très important, une approximation de l'erreur quadratique moyenne est donnée par

$$\mathrm{AMSE}(\hat{t}^d_{y,syn}) \simeq \mathbb{E}(\hat{t}^d_{y,syn} - \hat{t}^d_{y,dir})^2 - \mathrm{Var}(\hat{t}^d_{y,dir}),$$

où $\hat{t}^d_{y,dir}$ est un estimateur direct quelconque de $t^d_y$. Un estimateur de l'erreur quadratique moyenne est donné par

$$\widehat{\mathrm{MSE}}(\hat{t}^d_{y,syn}) = (\hat{t}^d_{y,syn} - \hat{t}^d_{y,dir})^2 - \widehat{\mathrm{Var}}(\hat{t}^d_{y,dir}).$$

Cet estimateur est approximativement sans biais mais très instable, i.e. pouvant prendre des valeurs négatives. Gonzalez & Wakesberg (1973) ont alors proposé un estimateur stable mais qui ne dépend plus spécifiquement du domaine

$$\widehat{\mathrm{MSE}}^{GW}(\hat{t}^d_{y,syn}) = \frac{1}{D}\sum_{d=1}^{D}(\hat{t}^d_{y,syn} - \hat{t}^d_{y,dir})^2 - \frac{1}{D}\sum_{d=1}^{D}\widehat{\mathrm{Var}}(\hat{t}^d_{y,dir}).$$

**Définition 1.14** *Une estimation au niveau d'un domaine est composite lorsqu'elle est le résultat d'une moyenne pondérée d'un estimateur direct et d'un estimateur synthétique.*

**Estimateur composite** Un estimateur de type composite du total $t^d_y$ s'écrit sous la forme

$$\hat{t}^d_{y,com} = \phi_d \hat{t}^d_{y,dir} + (1 - \phi_d)\hat{t}^d_{y,syn},$$

où le paramètre $\phi_d$ est un réel compris entre 0 et 1. Il existe plusieurs manières de choisir le paramètre $\phi_d, d = 1, \ldots, D$. Drew et al. (1982) ont, par exemple, proposé un paramètre $\phi_d$ qui dépend de la taille du domaine dans l'échantillon. Plus précisément, pour $\delta > 0$,

$$\phi_d = \begin{cases} 1 & \text{si } \hat{N}_d \geq \delta N_d, \\ \frac{\hat{N}_d}{\delta N_d} & \text{sinon.} \end{cases}$$

L'estimateur de type composite est optimal si le paramètre $\phi_d$ minimise l'erreur quadratique moyenne de l'estimateur. L'erreur quadratique moyenne de l'estimateur de type composite est de la forme

$$\begin{aligned} \mathrm{MSE}(\hat{t}^d_{y,com}) &= \phi_d^2 \mathrm{MSE}(\hat{t}^d_{y,dir}) + (1 - \phi_d)^2 \mathrm{MSE}(\hat{t}^d_{y,syn}) \\ &\quad + 2\phi_d(1 - \phi_d)\mathbb{E}(\hat{t}^d_{y,dir} - t^d_y)(\hat{t}^d_{y,syn} - t^d_y). \end{aligned}$$

Le paramètre $\phi_d^*$ optimal est approché par

$$\phi_d^* \simeq \left[1 + \frac{\text{MSE}(\hat{t}_{y,dir}^d)}{\text{MSE}(\hat{t}_{y,syn}^d)}\right]^{-1},$$

en supposant le terme $\mathbb{E}(\hat{t}_{y,dir}^d - t_y^d)(\hat{t}_{y,syn}^d - t_y^d)$ relativement négligeable. Le paramètre optimal $\phi_d^*$ dépend donc des vraies erreurs quadratiques moyennes $\text{MSE}(\hat{t}_{y,dir}^d)$ et $\text{MSE}(\hat{t}_{y,syn}^d)$. Un estimateur de ce paramètre peut être donné par

$$\hat{\phi}_d^* = \frac{\widehat{\text{MSE}}(\hat{t}_{y,syn}^d)}{(\hat{t}_{y,syn}^d - \hat{t}_{y,dir}^d)^2},$$

qui a cependant la particularité d'être très instable.

Une autre approche pour le choix du paramètre consiste donc à prendre un paramètre commun $\phi_d = \phi$ pour tout $d = 1, \ldots, D$, qui serait obtenu en minimisant la quantité $\sum_{d=1}^D \text{MSE}(\hat{t}_{y,com}^d)$ qui est la somme sur les domaines des erreurs quadratiques moyennes des estimateurs composites. Cette approche a été proposée par Purcell & Kish (1979). Un estimateur $\hat{\phi}^*$ de ce paramètre optimal commun $\phi^*$ est donné par

$$\begin{aligned}\hat{\phi}^* &= \frac{\sum_{d=1}^D \widehat{\text{MSE}}(\hat{t}_{y,syn}^d)}{\sum_{d=1}^D (\hat{t}_{y,syn}^d - \hat{t}_{y,dir}^d)^2}, \\ &= 1 - \frac{\sum_{d=1}^D \widehat{\text{Var}}(\hat{t}_{y,dir}^d)}{\sum_{d=1}^D (\hat{t}_{y,syn}^d - \hat{t}_{y,dir}^d)^2}.\end{aligned}$$

L'estimateur obtenu est alors stable mais ne dépend plus spécifiquement du domaine. Il est à noter que l'estimateur de type composite construit avec le nouveau paramètre $\hat{\phi}^*$ est alors de la même forme que l'estimateur de James & Stein (1961).

### 1.2.2 Méthodes utilisant l'approche basée sur un modèle

Les méthodes sur petits domaines basées explicitement sur un modèle reposent sur l'idée qu'un modèle relie la variable d'intérêt avec les variables auxiliaires. La variable d'intérêt ainsi que les variables auxiliaires ne sont donc que le résultat d'un processus aléatoire, celui du modèle. La source de variabilité n'est plus seulement liée au sondage, elle est aussi liée à l'inférence sous le modèle considéré.

Deux types de modèles existent pour l'estimation sur petits domaines. Le premier est une modélisation portant sur les grandeurs définies au niveau des domaines (modèle au niveau des domaines). De tels modèles sont essentiels lorsqu'aucune unité d'observation n'est disponible au niveau du domaine et lorsque le nombre de domaines est très grand. Le second type

est une modélisation portant sur les unités (modèle au niveau des unités). Des modèles utilisant ces deux niveaux peuvent ensuite être obtenus.

**Modèle au niveau du domaine**

L'approche la plus connue conçue au niveau du domaine est sans doute le modèle de Fay & Herriot (1979). Considérons toujours une partition de la population en $D$ domaines $U_d, d = 1, \ldots, D$, de tailles respectives $N_d, d =, \ldots, D$. L'objectif est maintenant de s'intéresser à une fonction $g(.)$ de la moyenne réelle de la variable d'intérêt. Comme dans cette approche, la variable d'intérêt est considérée comme la réalisation d'un modèle, nous noterons $\overline{Y}_d$ cette moyenne. Nous notons donc

$$\vartheta_d = g(\overline{Y}_d),$$

pour $d = 1, \ldots, D$, les paramètres d'intérêt. Par la suite, la fonction $g(.)$ représentera généralement la fonction *identité*.

Supposons maintenant que pour chaque paramètre $\vartheta_d$, nous avons le modèle

$$\xi : \vartheta_d = \mathbf{z}_d^\top \beta + b_d v_d, \text{ pour } d = 1, \ldots, D,$$

où $\mathbf{z}_d$ est un vecteur de covariables connu pour chaque domaine $U_d$, $\beta$ est un vecteur inconnu, $b_d$ est un réel connu et $v_d$ est une variable aléatoire centrée, i.e. $\mathbb{E}_\xi(v_d) = 0$ et de variance $\sigma_v^2$ inconnue, i.e. $\mathrm{Var}_\xi(v_d) = \sigma_v^2$. Les variables $v_d, d = 1, \ldots, D$ sont supposées mutuellement indépendantes.

Sur les données de l'enquête, on dispose d'un estimateur direct $\hat{\vartheta}_{d,dir}$ de $\vartheta_d$, ce qui permet d'écrire

$$\hat{\vartheta}_{d,dir} = \vartheta_d + e_d,$$

où les termes $e_d, d = 1, \ldots, D$ représentent les erreurs d'échantillonnage telle que $\mathbb{E}(e_d|\vartheta_d) = 0$ et $\mathrm{Var}(e_d|\vartheta_d) = \psi_d$ supposées connues. Les termes $e_d, d = 1, \ldots, D$ sont supposés indépendants entre eux. Les termes $e_d$ et $v_d$ sont également supposés indépendants deux à deux. On obtient donc finalement le modèle combiné de Fay & Herriot (1979)

$$\hat{\vartheta}_{d,dir} = \mathbf{z}_d^\top \beta + b_d v_d + e_d, \text{ pour } d = 1, \ldots, D.$$

**Estimateur BLUP** Le meilleur estimateur linéaire sans biais de $\vartheta_d$ qui minimise l'erreur quadratique moyenne (BLUP), est noté $\tilde{\vartheta}_d$ et est alors donné par

$$\tilde{\vartheta}_d = \mathbf{z}_d^\top \tilde{\beta} + \gamma_d(\hat{\vartheta}_{d,dir} - \mathbf{z}_d^\top \tilde{\beta}),$$

avec

$$\gamma_d = \frac{b_d^2 \sigma_v^2}{\psi_d + b_d^2 \sigma_v^2},$$

et

$$\tilde{\beta} = \left( \sum_{d=1}^{D} \gamma_d \mathbf{z}_d \mathbf{z}_d^\top \right)^{-1} \sum_{d=1}^{D} \gamma_d \mathbf{z}_d \hat{\vartheta}_{d,dir}.$$

L'estimateur $\tilde{\vartheta}_d$ peut être réécrit sous la forme d'un estimateur composite

$$\tilde{\vartheta}_d = \gamma_d \hat{\vartheta}_{d,dir} + (1 - \gamma_d) \mathbf{z}_d^\top \tilde{\beta}.$$

Lorsque la variance d'échantillonnage $\psi_d$ est relativement petit, l'estimateur $\tilde{\vartheta}_d$ donnera plus d'importance à l'estimateur direct. Et inversement, lorsque $\psi_d$ augmente ou lorsque $\sigma_v^2$ diminue, il donnera plus d'importance à l'estimateur synthétique.

Il convient cependant de noter que l'estimateur $\tilde{\vartheta}_d$ dépend de $\sigma_v^2$. Le meilleur estimateur empirique linéaire sans biais de $\vartheta_d$ qui minimise l'erreur quadratique moyenne (Eblup) s'obtient en remplaçant $\sigma_v^2$ par $\hat{\sigma}_v^2$ dans l'expression de $\tilde{\vartheta}_d$

$$\hat{\vartheta}_d = \tilde{\vartheta}_d(\hat{\sigma}_v^2).$$

Fay & Herriot (1979) ont proposé un estimateur de la variance $\sigma_v^2$ en résolvant itérativement par la méthode des moments

$$\sum_{d=1}^{D} \frac{(\hat{\vartheta}_{d,dir} - \mathbf{z}_d^\top \tilde{\beta})^2}{\psi_d + b_d^2 \sigma_v^2} = m - J,$$

$J$ étant le nombre de paramètre du modèle, sans aucune hypothèse de normalité. D'autres estimateurs de la variance $\sigma_v^2$ peuvent être obtenus en utilisant la méthode de maximum de vraisemblance ou la méthode de maximum de vraisemblance restreinte en supposant que les $\hat{\vartheta}_{d,dir}, d = 1, \ldots, D$ suivent la distribution d'une loi normale d'espérance $\mathbf{z}_d^\top \beta$ et de variance $\psi_d + b_d^2 \sigma_v^2$.

**Estimateur Bayésien** Considérant l'hypothèse de normalité dans les modèles de Fay & Herriot (1979), l'estimateur Bayésien $\tilde{\vartheta}_d^B$ de $\vartheta_d$ est donné par

$$\tilde{\vartheta}_d^B(\beta, \sigma_v^2) = \mathbb{E}_\xi(\vartheta_d | \hat{\vartheta}_{d,dir}) = \gamma_d \hat{\vartheta}_{d,dir} + (1 - \gamma_d) \mathbf{z}_d^\top \beta.$$

L'estimateur Bayésien $\tilde{\vartheta}_d^B$ est sans biais car

$$\mathbb{E}_\xi(\tilde{\vartheta}_d^B) = \mathbb{E}_{\hat{\vartheta}_{d,dir}} \mathbb{E}_{\vartheta_d | \hat{\vartheta}_{d,dir}}(\vartheta_d) = \vartheta_d.$$

L'estimateur empirique Bayésien $\hat{\vartheta}_d^{EB}$ de $\tilde{\vartheta}_d^B$ s'obtient en remplaçant $\beta$ et $\sigma_v^2$ par $\hat{\beta}$ et $\hat{\sigma}_v^2$

$$\hat{\vartheta}_d^{EB} = \tilde{\vartheta}_d^B(\hat{\beta}, \hat{\sigma}_v^2) = \hat{\vartheta}_d,$$

qui a ainsi la particularité de coïncider avec l'estimateur Eblup.

L'erreur quadratique moyenne de l'estimateur EBLUP peut être approchée en utilisant la méthode de linéarisation de Taylor. Sous l'hypothèse de normalité, on a

$$\begin{aligned} \mathrm{MSE}_{\xi}(\hat{\vartheta}_d) &= \mathbb{E}_{\xi}(\hat{\vartheta}_d - \vartheta_d)^2 \\ &\simeq g_{1d}(\sigma_v^2) + g_{2d}(\sigma_v^2) + g_{3d}(\sigma_v^2), \end{aligned}$$

avec

$$\begin{aligned} g_{1d}(\sigma_v^2) &= \gamma_d \psi_d, \\ g_{2d}(\sigma_v^2) &= (1-\gamma_d)^2 \mathbf{z}_d^{\top} \left( \sum_{d=1}^{D} \frac{\mathbf{z}_d \mathbf{z}_d^{\top}}{\psi_d + b_d^2 \sigma_v^2} \right)^{-1} \mathbf{z}_d, \\ g_{3d}(\sigma_v^2) &= \psi_d^3 b_d^4 (\psi_d + b_d^2 \sigma_v^2)^{-3} \mathrm{AVar}(\hat{\sigma}_v^2), \end{aligned}$$

où $\mathrm{AVar}(\hat{\sigma}_v^2)$ représente une approximation de la variance de $\hat{\sigma}_v^2$. Lorsque l'estimateur $\hat{\sigma}_v^2$ est obtenu par la méthode de maximum de vraisemblance restreinte, un estimateur de l'erreur quadratique moyenne *quasi* sans biais est donné par

$$\widehat{\mathrm{MSE}}_{\xi}(\hat{\vartheta}_d) = g_{1d}(\hat{\sigma}_v^2) + g_{2d}(\hat{\sigma}_v^2) + 2g_{3d}(\hat{\sigma}_v^2)$$

où

$$\mathbb{E}_{\xi}[\widehat{\mathrm{MSE}}_{\xi}(\hat{\vartheta}_d)] = \mathrm{MSE}_{\xi}(\hat{\vartheta}_d) + o(\frac{1}{D}).$$

Lorsque l'estimateur $\hat{\sigma}_v^2$ est obtenu par la méthode de maximum de vraisemblance ou par la méthode de Fay & Herriot (1979), un terme qui accompagne le biais de $\hat{\sigma}_v^2$ est ajouté (Rao, 2003, p. 129).

**Modèle au niveau des unités**

Dans cette partie, la modélisation se porte directement sur les unités d'observation. Dans la même hypothèse d'une partition de la population en $D$ domaines $U_d, d = 1, \ldots, D$, de tailles respectives $N_d, d =, \ldots, D$, soit $y_{dk}$ la valeur de la variable d'intérêt pour l'unité $k$ du domaine $U_d$. La variable d'intérêt est supposée suivre un modèle de régression linéaire à erreurs emboîtées (Henderson, 1975; Battese et al., 1988; Prasad & Rao, 1990; You & Rao, 2002; Rao, 2003) tel que

$$\xi : y_{dk} = \mathbf{x}_{dk}^{\top}\beta + v_d + \varepsilon_{dk}, \text{ pour } k \in U_d \text{ et } d = 1, \ldots, D,$$

où $\mathbf{x}_{dk}$ représentent les variables auxiliaires disponibles pour l'unité $k$ du domaine $U_d$, $\beta$ est un vecteur inconnu, $v_d$ est une variable aléatoire suivant une distribution de la loi normale centrée et de variance $\sigma_v^2$ inconnue, et $\varepsilon_{dk}$ est un terme aléatoire suivant une distribution de la loi normale centrée et de variance $\kappa_{dk}^2 \sigma_{\varepsilon}^2$, le terme $\kappa_{dk}^2$ modélisant une éventuelle hétéroscédasticité. Les termes $v_d$ et $\varepsilon_{dk}$ sont supposés indépendants,

et indépendants deux à deux. Le modèle $\xi$ est supposé vrai pour toutes les unités de la population $U$ ainsi que toutes les unités de l'échantillon $S$ qui est aussi partitionné en $D$ domaines,$S_d, d = 1, \ldots, D$.

Dans l'approche basée sur un modèle de régression linéaire à erreurs emboîtées, deux formes du meilleur prédicteur linéaire sans biais (BLUP) co-existent : l'estimateur proposé par Royall (1970) et l'estimateur proposé par Henderson et al. (1959) et par la suite formalisé par Henderson (1975). Ces deux estimateurs ne sont pas égaux et sont les solutions de différentes procédures d'optimisation (voir, par exemple, Guggemos & Tillé, 2009).

**Estimateur BLUP de Royall (1970)** Matriciellement, le modèle $\xi$ peut être réécrit comme suit

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \boldsymbol{\varepsilon}$$

tel que

$$\mathbb{E}_\xi(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}\ ,\ \mathrm{Var}_\xi(\mathbf{y}) = \mathbf{V} = \sigma_v^2\mathbf{Z}\mathbf{Z}^\top + \sigma_\varepsilon^2 \mathrm{diag}_{k=1,\ldots,N_d}^{d=1,\ldots,D}(\kappa_{dk}^2).$$

Par décomposition de la population en deux parties (partie incluse dans l'échantillon et partie hors de l'échantillon), on obtient

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{pmatrix}, \mathbf{y} = \begin{pmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{pmatrix}.$$

Le paramètre d'intérêt linéaire devient alors

$$\delta = \mathbf{a}^\top\mathbf{y} = \mathbf{a}_s^\top\mathbf{y}_s + \mathbf{a}_r^\top\mathbf{y}_r.$$

Le meilleur estimateur linéaire sans biais de $\delta$ qui minimise l'erreur quadratique moyenne (BLUP) est noté $\tilde{\delta}^R$ et est alors donné par

$$\tilde{\delta}^R = \mathbf{a}_s^\top\mathbf{y}_s + \mathbf{a}_r^\top[\mathbf{X}_r\tilde{\boldsymbol{\beta}} + \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s\tilde{\boldsymbol{\beta}})],$$

avec

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}_s^\top\mathbf{V}_{ss}^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s^\top\mathbf{V}_{ss}^{-1}\mathbf{y}_s.$$

Dans le cas où le paramètre d'intérêt est le total de la variable $y$ au niveau du domaine $U_d$, i.e. $\delta = \delta_d = t_y^d$, on obtient

$$\tilde{\delta}_d^R = \sum_{k\in S_d} y_{dk} + \sum_{k\in U_d\setminus S_d}[\mathbf{x}_{dk}^\top\tilde{\beta} + \gamma_d(\overline{y}_{da} - \overline{\mathbf{x}}_{da}^\top\tilde{\beta})],$$

avec

$$\gamma_d = \frac{\sigma_v^2}{\sigma_v^2 + \frac{\sigma_\varepsilon^2}{a_{d.}}}\ ,\ a_{d.} = \sum_{k\in S_d} a_{dk} = \sum_{k\in S_d}\kappa_{dk}^{-2}$$

et

$$\overline{y}_{da} = \sum_{k\in S_d}\frac{a_{dk}}{a_{d.}}y_{dk}\ ,\ \overline{\mathbf{x}}_{da} = \sum_{k\in S_d}\frac{a_{dk}}{a_{d.}}\mathbf{x}_{dk}.$$

L'estimateur BLUP de Royall (1970) $\tilde{\delta}_d^R$ du total $t_y^d$ dépend bien évidemment des variances $\sigma_v^2$ et $\sigma_\varepsilon^2$. L'estimateur EBLUP de Royall (1970) noté $\hat{\delta}_d^R$ s'obtient en remplaçant les variances $\sigma_v^2$ et $\sigma_\varepsilon^2$ par leurs estimateurs $\hat{\sigma}_v^2$ et $\hat{\sigma}_\varepsilon^2$. Ces estimateurs des variances peuvent être obtenus par la méthode des moments, par la méthode du maximum de vraisemblance ou par la méthode du maximum de vraisemblance restreinte. Finalement, on obtient donc

$$\hat{\delta}_d^R = \tilde{\delta}_d^R(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2).$$

**Estimateur BLUP de Henderson et al. (1959); Henderson (1975)** Il convient de rappeler que le modèle de régression linéaire à erreurs emboîtées $\xi$ est valable au niveau de tout domaine, que ce soit dans la population ou dans l'échantillon. Considérons donc le modèle au niveau de l'échantillon comme suit

$$y_{dk} = \mathbf{x}_{dk}^\top \beta + v_d + \varepsilon_{dk}, \text{ pour } k \in S_d \text{ et } d = 1, \ldots, D.$$

L'objectif est maintenant d'estimer la moyenne

$$\mu_d = \overline{\mathbf{X}}_d^\top \beta + v_d$$

telle que le vecteur des moyennes des variables auxiliaires

$$\overline{\mathbf{X}}_d = \frac{\mathbf{t}_x^d}{N_d}$$

est toujours supposé connu. Le meilleur estimateur linéaire sans biais de $\mu_d$ qui minimise l'erreur quadratique moyenne (BLUP) est noté $\tilde{\mu}_d^H$ et est alors donné par

$$\tilde{\mu}_d^H = \overline{\mathbf{X}}_d^\top \tilde{\beta} + \gamma_d \left( \overline{y}_{da} - \overline{\mathbf{x}}_{da}^\top \tilde{\beta} \right)$$

avec

$$\overline{y}_{da} = \sum_{k \in S_d} \frac{a_{dk}}{a_{d.}} y_{dk} \text{ , } \overline{\mathbf{x}}_{da} = \sum_{k \in S_d} \frac{a_{dk}}{a_{d.}} \mathbf{x}_{dk} \text{ ,}$$

$$\gamma_d = \frac{\sigma_v^2}{\sigma_v^2 + \frac{\sigma_\varepsilon^2}{a_{d.}}} \text{ , } a_{d.} = \sum_{k \in S_d} a_{dk} = \sum_{k \in S_d} \kappa_{dk}^{-2},$$

et

$$\tilde{\beta} = \left( \sum_{k \in S_d} a_{dk} \mathbf{x}_{dk} \mathbf{x}_{dk}^\top - \gamma_d a_{d.} \overline{\mathbf{x}}_{da} \overline{\mathbf{x}}_{da}^\top \right)^{-1} \left( \sum_{k \in S_d} a_{dk} \mathbf{x}_{dk} y_{dk}^\top - \gamma_d a_{d.} \overline{\mathbf{x}}_{da} \overline{y}_{da}^\top \right).$$

L'estimateur BLUP de Henderson (1975) $\tilde{\mu}_d^H$ de la moyenne $\mu_d$ peut se réécrire sous la forme d'un estimateur de type composite

$$\tilde{\mu}_d^H = \gamma_d [\overline{y}_{da} + (\overline{\mathbf{X}}_d - \overline{\mathbf{x}}_{da})^\top \tilde{\beta}] + (1 - \gamma_d) \overline{\mathbf{X}}_d^\top \tilde{\beta}.$$

De même, l'estimateur EBLUP de Henderson (1975) noté $\hat{\mu}_d^H$ s'obtient en remplaçant les variances $\sigma_v^2$ et $\sigma_\varepsilon^2$ par leurs estimateurs $\hat{\sigma}_v^2$ et $\hat{\sigma}_\varepsilon^2$ qui sont

par exemple obtenus par la méthode des moindres carrés et la méthode des moments. On obtient donc

$$\hat{\mu}_d^H = \tilde{\mu}_d^H(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2).$$

L'erreur quadratique moyenne de l'estimateur EBLUP $\hat{\mu}_d^H$ peut être approchée par

$$\text{MSE}_\xi(\hat{\mu}_d^H) \simeq g_{1d}(\sigma_v^2, \sigma_\varepsilon^2) + g_{2d}(\sigma_v^2, \sigma_\varepsilon^2) + g_{3d}(\sigma_v^2, \sigma_\varepsilon^2),$$

avec

$$\begin{aligned}
g_{1d}(\sigma_v^2, \sigma_\varepsilon^2) &= \gamma_d \frac{\sigma_\varepsilon^2}{a_{d.}}, \\
g_{2d}(\sigma_v^2, \sigma_\varepsilon^2) &= (\overline{\mathbf{X}}_d - \gamma_d \overline{\mathbf{x}}_{da})^\top \left( \sum_{k \in S_d} a_{dk} \mathbf{x}_{dk} \mathbf{x}_{dk}^\top - \gamma_d a_{d.} \overline{\mathbf{x}}_{da} \overline{\mathbf{x}}_{da}^\top \right)^{-1} (\overline{\mathbf{X}}_d - \gamma_d \overline{\mathbf{x}}_{da}), \\
g_{3d}(\sigma_v^2, \sigma_\varepsilon^2) &= a_{d.}^{-2} \left( \sigma_v^2 + \frac{\sigma_\varepsilon^2}{a_{d.}} \right)^{-3} \Big[ \sigma_\varepsilon^4 \text{AVar}(\sigma_v^2) + \sigma_v^2 \text{AVar}(\sigma_\varepsilon^2) \\
&\quad - 2\sigma_\varepsilon^2 \sigma_v^2 \text{ACov}(\sigma_\varepsilon^2, \sigma_v^2) \Big]
\end{aligned}$$

où AVar(.) et ACov(.) représentent les approximations de variances et de covariance. Un estimateur de l'erreur quadratique moyenne de $\hat{\mu}_d^H$ est alors donné par

$$\widehat{\text{MSE}}_\xi(\hat{\mu}_d^H) = g_{1d}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) + g_{2d}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) + 2g_{3d}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2).$$

**Estimateur pseudo-EBLUP** Dérivé de l'estimateur BLUP de Henderson (1975), Prasad & Rao (1999) et You & Rao (2002) ont proposé un estimateur utilisant des poids de sondages $w_{dk}, k \in S_d, d = 1, \ldots, D$ tels que pour tout domaine $d$, $\sum_{k \in S_d} w_{dk} = N_d$. Grâce à un habile changement de variables, l'estimateur proposé par You & Rao (2002) a la particularité de respecter la propriété de cohérence, selon laquelle, pour une variable d'intérêt donnée, la somme des estimations des totaux de la variable au niveau des domaines coïncide avec l'estimation directe du total au niveau de la population. Plus précisément, l'estimateur BLUP de la moyenne $\mu_d$ proposé par You & Rao (2002) peut être généralisé comme suit

$$\tilde{\mu}_{dw}^H = \overline{\mathbf{X}}_d^\top \tilde{\beta}_w + \gamma_{dw} \left( \overline{y}_{daw} - \overline{\mathbf{x}}_{daw}^\top \tilde{\beta}_w \right),$$

où

$$\gamma_{dw} = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\varepsilon^2 \delta_{dw}}, \; \delta_{dw} = \frac{\sum_{k \in S_d} w_{dk}^2 a_{dk}}{(\sum_{k \in S_d} w_{dk} a_{dk})^2}, \; a_{dk} = \kappa_{dk}^{-2},$$

avec

$$\overline{y}_{daw} = \frac{\sum_{k \in S_d} w_{dk} a_{dk} y_{dk}}{\sum_{k \in S_d} w_{dk} a_{dk}}, \; \overline{\mathbf{x}}_{daw} = \frac{\sum_{k \in S_d} w_{dk} a_{dk} \mathbf{x}_{dk}}{\sum_{k \in S_d} w_{dk} a_{dk}},$$

et

$\tilde{\beta}_w = \left[\sum_{d=1}^{D}\sum_{k\in S_d} w_{dk}a_{dk}\mathbf{x}_{dk}(\mathbf{x}_{dk}-\gamma_{dw}\bar{\mathbf{x}}_{daw})^\top\right]^{-1}\sum_{d=1}^{D}\sum_{k\in S_d} w_{dk}a_{dk}(\mathbf{x}_{dk}-\gamma_{dw}\bar{\mathbf{x}}_{daw})y_{dk}$.

En remplaçant les variances $\sigma_\varepsilon^2$ et $\sigma_v^2$ par leurs estimateurs $\hat{\sigma}_\varepsilon^2$ et $\hat{\sigma}_v^2$ dans les formules de $\gamma_{dw}$ et $\tilde{\beta}_w$, l'estimateur pseudo-EBLUP de $\mu_d$ est alors donné par

$$\hat{\mu}_{dw}^H = \tilde{\mu}_{dw}^H(\hat{\sigma}_\varepsilon^2, \hat{\sigma}_v^2).$$

Sous la condition que le modèle $\xi$ inclut l'ordonnée à l'origine non nulle, You & Rao (2002) ont alors montré que

$$\sum_{d=1}^{D} N_d\hat{\mu}_{dw}^H = \sum_{d=1}^{D}\sum_{k\in S_d} w_{dk}y_{dk} + (\sum_{d=1}^{D}\mathbf{t}_x^d - \sum_{d=1}^{D}\sum_{k\in S_d} w_{dk}\mathbf{x}_{dk})\hat{\beta}_w,$$

i.e. la somme des estimateurs des totaux des domaines coïncide avec l'estimateur direct par la régression.

Un estimateur de l'erreur quadratique moyenne de l'estimateur pseudo-EBLUP $\hat{\mu}_{dw}^H$ basé sur celui proposé par You & Rao (2002) peut être donné par

$$\widehat{\mathrm{MSE}}(\hat{\mu}_{dw}^H) = g_{1dw}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) + g_{2dw}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) + 2g_{3dw}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2)$$

avec

$$\begin{aligned}
g_{1dw}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) &= \hat{\gamma}_{dw}\delta_{dw}\hat{\sigma}_\varepsilon^2,\\
g_{2dw}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) &= (\overline{\mathbf{X}}_d - \hat{\gamma}_{dw}\bar{\mathbf{x}}_{daw})^\top\Phi_w(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2)(\overline{\mathbf{X}}_d - \hat{\gamma}_{dw}\bar{\mathbf{x}}_{daw}),\\
g_{3dw}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) &= \frac{\hat{\gamma}_{dw}(1-\hat{\gamma}_{dw})^2}{\hat{\sigma}_v^2\hat{\sigma}_\varepsilon^2}\left[\hat{\sigma}_\varepsilon^4\mathrm{AVar}(\hat{\sigma}_v^2) + \hat{\sigma}_v^4\mathrm{AVar}(\hat{\sigma}_\varepsilon^2)\right.\\
&\quad \left.-2\hat{\sigma}_\varepsilon^2\hat{\sigma}_v^2\mathrm{ACov}(\hat{\sigma}_\varepsilon^2, \hat{\sigma}_v^2)\right],
\end{aligned}$$

où en posant $\mathbf{z}_{dk} = w_{dk}a_{dk}(\mathbf{x}_{dk} - \hat{\gamma}_{dw}\bar{\mathbf{x}}_{daw})$,

$$\begin{aligned}
\Phi_w(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) &= \hat{\sigma}_\varepsilon^2\left(\sum_d\sum_k \mathbf{x}_{dk}\mathbf{z}_{dk}^\top\right)^{-1}\left(\sum_d\sum_k \mathbf{z}_{dk}\mathbf{z}_{dk}^\top\right)\left[\left(\sum_d\sum_k \mathbf{x}_{dk}\mathbf{z}_{dk}^\top\right)^{-1}\right]^\top\\
&\quad +\hat{\sigma}_v^2\left(\sum_d\sum_k \mathbf{x}_{dk}\mathbf{z}_{dk}^\top\right)^{-1}\left[\sum_d\left(\sum_k \mathbf{z}_{dk}\right)\left(\sum_k \mathbf{z}_{dk}\right)^\top\right]\\
&\quad \left[\left(\sum_d\sum_k \mathbf{x}_{dk}\mathbf{z}_{dk}^\top\right)^{-1}\right]^\top.
\end{aligned}$$

L'estimateur $\widehat{\mathrm{MSE}}(\hat{\mu}_{dw}^H)$ de l'erreur quadratique moyenne est *quasi* sans biais tel que

$$\mathbb{E}_\xi[\widehat{\mathrm{MSE}}_\xi(\hat{\mu}_{dw}^H)] = \mathrm{MSE}_\xi(\hat{\mu}_{dw}^H) + o(\frac{1}{D}).$$

### 1.2.3 Discussion

Les méthodes d'estimation sur petits domaines présentées montrent clairement l'importance de l'information auxiliaire disponible au niveau de chaque domaine. L'utilisation de cette information constitue *"l'essence même de toutes méthodes pour petits domaines"* (Rao, 2003, p. xviii).

L'inconvénient des méthodes d'estimation directe basées sur le plan de sondage est que leur précision est faible lorsque la taille d'un domaine est petite. Une solution à ce problème consiste à emprunter de la "force" au niveau des autres domaines. Cette idée a été essentiellement développée dans le cadre de l'approche basée sur un modèle. Il est intéressant de remarquer que les "bons" estimateurs développés par cette approche sont de type composite. Les unités statistiques ont deux types de contribution dans l'estimation : d'une part, une "auto-contribution" ou contribution de l'unité à son propre domaine, et une "extra-contribution" ou contribution de l'unité à un domaine autre que le sien, d'autre part. Dans cette thèse, des méthodes d'estimation exploitant la même idée sont proposées mais avec une approche basée sur le plan de sondage.

# 2

# Optimisation du redressement d'une enquête : application à un sous-échantillon de l'Enquête Nationale sur les Transports et les Déplacements 2007-2008

**Résumé** Une optimisation du redressement par calage d'un sous-échantillon d'une enquête est proposée. Disposant de nombreuses variables auxiliaires connues au niveau du sous-échantillon ou domaine, le but de ce travail consiste à éviter le phénomène dit de "sur-calage" : des difficultés à satisfaire les équations de calage et une saturation du nombre de degré de libertés ayant pour conséquences une explosion des poids et une augmentation de la variance, voire une impossibilité de calculer les poids de calage. L'exercice consiste à choisir judicieusement les variables auxiliaires jugées pertinentes afin d'améliorer la qualité des estimations au niveau du domaine. La précision des estimations est quantifiée par le calcul de la variance des estimateurs calés ainsi obtenus. L'estimation des variances est basée sur la méthode de la linéarisation et sur les techniques des résidus.[1]

**Mots clés** redressement, enquête, sous-échantillon, calage, variance

## 2.1 Introduction

La qualité des estimations issues d'une enquête par sondage peut être améliorée en présence d'information auxiliaire (voir par exemple, Tillé, 1992). Le calage sur marges est une méthode de redressement d'enquête efficace lorsque la taille de l'échantillon est suffisamment grande (Deville

[1]Ce chapitre est une version reprise d'un article, co-écrit avec Jimmy Armoogum, soumis pour publication dans une revue scientifique internationale.

& Särndal, 1992). En disposant d'un grand nombre de variables auxiliaires et en se restreignant à un sous-échantillon, le redressement par calage peut conduire à des instabilités des poids provoquant ainsi une diminution de la précision des estimations (voir, par exemple, Chauvet & Goga, 2012). Cet article a pour but de mener une discussion sur le choix des variables auxiliaires à utiliser lors d'un redressement au niveau d'un sous-échantillon. La variance d'un estimateur est un excellent outil pour quantifier sa précision. L'obtention de la variance minimale pour une variable d'intérêt donnée dépend des variables auxiliaires choisies. Les variables auxiliaires minimisant la variance peuvent donc être différentes d'une variable d'intérêt à une autre. En considérant plusieurs variables d'intérêt sur un même sujet, pas forcément très corrélées, nous proposons une méthode pour sélectionner les variables auxiliaires qui permettent d'établir un système de pondération unique pour différentes variables d'intérêt d'un même thème.

Un petit rappel sur le principe du calage sur marges est donné dans la Section 2.2. Dans la Section 2.3, nous donnons une présentation de l'Enquête Nationale sur les Transports et les Déplacements 2007-2008. La Section 2.4 présente la méthode proposée afin de sélectionner les variables auxiliaires nécessaires qui maximisent la précision des estimations. La Section 2.5 ouvre une discussion sur les résultats obtenus pour ainsi aboutir à une conclusion dans la Section 2.6.

## 2.2 Principe du calage sur marges et propriétés

L'idée générale de la méthode d'estimation par calage a été développée et formalisée par Deville & Särndal (1992), bien que de nombreux anciens travaux utilisaient déjà les méthodes d'ajustements de tableaux à des marges connues (Deming & Stephan, 1940; Lemel, 1976; Madre, 1979, 1980; Stephan, 1942). Selon Roux & Armoogum (2008, 2010), les techniques de calage sur marges "*consistent à faire coïncider les marges de quelques variables de l'échantillon à celles de la population en modifiant la pondération. Lorsque les variables auxiliaires sont qualitatives, cette approche ne nécessite pas la connaissance dans la population du croisement de ces variables auxiliaires.*"

### 2.2.1 Technique du calage

Soit une population finie $U = \{1, \ldots, k, \ldots, N\}$ dans laquelle un échantillon $s$ est tiré selon un plan de sondage donné $p(.)$. La quantité $p(s)$ représente la probabilité qu'un échantillon aléatoire $S$ prenne comme valeur l'échantillon $s$, i.e. $Pr(S = s) = p(s)$. La probabilité d'inclusion d'ordre 1 de l'unité $k$ dans l'échantillon est notée $\pi_k$. De même, la probabilité d'inclusion d'ordre 2 des unités $k$ et $l$ dans l'échantillon

est notée $\pi_{kl}$. Ces probabilités sont supposées strictement positives. Soit $x_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kJ})^\top$ un vecteur de caractères auxiliaires de l'unité $k$. Le vecteur des totaux de $\mathbf{x}$ dans la population, noté $\mathbf{t}_x$, est supposé connu. Soit $y_k$ la valeur de la variable d'intérêt pour l'unité $k$. L'objectif est d'estimer le total de la variable d'intérêt $\mathbf{y}$.

$$t_y = \sum_{k \in U} y_k. \tag{2.1}$$

En considérant l'information auxiliaire caractérisée par le vecteur $\mathbf{t}_x$ de totaux connus, l'estimateur par calage du total de la variable $\mathbf{y}$ s'écrit comme

$$\hat{t}_{y,w} = \sum_{k \in S} w_k y_k. \tag{2.2}$$

Évidemment, les poids $w_k$ dépendent de l'échantillon $s$ et satisfont l'équation de calage :

$$\sum_{k \in S} w_k \mathbf{x}_k^\top = \mathbf{t}_x. \tag{2.3}$$

où les poids $w_k$ doivent être proches des poids $d_k = 1/\pi_k$.

La proximité entre les poids $w_k$ et $d_k$ est définie en utilisant une pseudo-distance notée $G_k(.,.)$ supposée définie positive, dérivable et strictement convexe par rapport à $w_k$. Les poids $w_k$ sont obtenus en minimisant la quantité

$$\sum_{k \in S} \frac{G_k(w_k, d_k)}{q_k} \tag{2.4}$$

sous la contrainte de l'équation de calage (2.3). Les $1/q_k$ sont des coefficients de pondération qui déterminent l'importance de chaque unité dans le calcul de la distance.

Plusieurs distances peuvent être utilisées et sont discutées par Deville & Särndal (1992). En général, les poids $w_k$ s'obtiennent en résolvant en $\boldsymbol{\lambda}$, au moyen de la méthode de Newton, le système d'équation

$$\mathbf{t}_x = \sum_{k \in S} d_k \mathbf{x}_k^\top F(q_k \mathbf{x}_k \boldsymbol{\lambda}), \tag{2.5}$$

$\boldsymbol{\lambda}$ représentant le vecteur des $J$ multiplicateurs de Lagrange. Finalement,

$$w_k = d_k F(q_k \mathbf{x}_k \boldsymbol{\lambda}), \tag{2.6}$$

$F(.)$ représentant l'inverse de la fonction $g_k(w_k, d_k)$ qui est la dérivée de $G_k(w_k, d_k)$ par rapport à $w_k$.

Par la suite, nous utiliserons toujours la méthode logistique afin de

n'obtenir des poids ni trop élevés, ni négatifs. Dans ce cas, en considérant deux bornes strictement positives $L$ et $H$,

$$G_k(w_k, d_k) = \begin{cases} (a_k \log \frac{a_k}{1-L} + b_k \log b_k H - 1)\frac{1}{A} & \text{si } Ld_k < w_k < Hd_k \\ \infty & \text{sinon,} \end{cases} \tag{2.7}$$

où $a_k = \frac{w_k}{d_k} - L$, $b_k = H - \frac{w_k}{d_k}$ et $A = \frac{H-L}{(1-L)(H-1)}$.

### 2.2.2 Estimation de la variance d'un estimateur calé

L'estimation de la variance d'un estimateur par calage peut s'obtenir par la technique de linéarisation (voir par exemple, Tillé, 2001). Deville & Särndal (1992); Deville et al. (1993) ont montré que

$$\text{AVar}(\hat{t}_{y,w}) \simeq \text{Var}(\hat{t_E}) = \sum_{k\in U}\sum_{l\in U} \frac{E_k}{\pi_k}\frac{E_l}{\pi_l}\Delta_{kl}, \tag{2.8}$$

où $E_k = y_k - \mathbf{x}_k^\top \mathbf{B}$ donne les résidus de la régression de $\mathbf{y}$ sur le jeu des variables auxiliaires $\mathbf{x}$ au niveau de la population.

Un estimateur de la variance est alors donné par

$$\widehat{\text{Var}}(\hat{t}_{y,w}) = \sum_{k\in S}\sum_{l\in S} \frac{\Delta_{kl}}{\pi_{kl}} w_k e_k w_l e_l, \tag{2.9}$$

où $e_k = y_k - \mathbf{x}_k^\top \hat{\mathbf{B}}_s$ donne les résidus de la régression pondérée de $\mathbf{y}$ sur le jeu des variables auxiliaires $\mathbf{x}$ au niveau de l'échantillon.

Puisque l'estimateur de la variance se calcule à partir des résidus de la régression $w$-pondérée de $\mathbf{y}$ sur le jeu des variables auxiliaires $\mathbf{x}$, il est plus petit que la variance de l'estimateur de Horvitz & Thompson (1952). De plus, pour minimiser cette variance, il faut minimiser la somme du produit des poids de sondage et des résidus. L'introduction de l'information auxiliaire dans le calage permet de diminuer les résidus mais augmente aussi la dispersion des poids. Il faut donc choisir judicieusement les variables auxiliaires qui permettent de diminuer la variance lors du calage pour le redressement d'une enquête.

## 2.3 L'Enquête Nationale sur les Transports et les Déplacements 2007-2008

Dans cette Section, la présentation de l'ENTD 2007-2008 reprend celle[2] donnée par l'INSEE sur son site Internet. La description du redressement

[2]Le lien permanent de la présentation est donné par http://www.insee.fr/fr/methodes/default.asp?page=definitions/enq-transports-deplacements.htm.

de l'ENTD 2007-2008 est reprend largement les notes méthodologiques de Roux & Armoogum (2008, 2010).

### 2.3.1 Présentation de l'ENTD

L'INSEE présente l'ENTD 2007-2008 comme suit :

*"Tous les dix ans environ, le ministère chargé des Transports, l'INSEE et l'Institut National de Recherche sur les Transports et leur Sécurité*[3] *(INRETS) conduisent une Enquête Nationale sur les Transports (ENTD). L'ENTD 2007-2008 succède à celle de 1993-1994 et les précédentes enquêtes datent de 1966-67, 1973-74 et 1981-82. L'objectif de ces enquêtes est la connaissance des déplacements des ménages résidant en France et de leur usage des moyens de transport tant collectifs qu'individuels. Elle permet d'avoir une vision globale et cohérente de la mobilité et d'analyser le parc de véhicules dont disposent les ménages et de leur usage. Elle permet aussi de répondre aux questions sur les trafics interrégionaux et internationaux dont les enjeux sont très importants en matière d'investissements et de mesurer les distances parcourues dont la connaissance est indispensable pour appréhender les problématiques environnementales. Par rapprochement avec les résultats des enquêtes précédentes, elle permet des comparaisons dans le temps et dans l'espace."*

### 2.3.2 Redressement de l'ENTD

Le recensement de la population de 1999 est une source d'information auxiliaire complète permettant de redresser l'ENTD. Roux & Armoogum (2008, 2010, 2011); Armoogum & Roux (2012) ont mis en évidence les variables auxiliaires qui permettent d'expliquer le mécanisme de réponse au moyen d'un modèle logistique et ont ainsi retenu les variables suivantes :

- **Type de bâtiment** (logement collectif ; maison),
- **Nombre de pièces du logement** (studio ou chambre ; 2-3 pièces ; 4-5 pièces ; 6 pièces et plus),
- **Zone de résidence** (commune rurale ; unité urbaine de moins de 20000 habitants ; unité urbaine de 20000 à 99999 habitants ; unité urbaine de plus de 100000 habitants ; unité urbaine de Paris),
- **Motorisation du ménage** au recensement de 1999 (0 voiture, 1 voiture ; 2 voitures et plus),
- **Âge de la personne de référence** au recensement de 1999 (15-34 ans ; 35-49 ans ; 50-64 ans ; 65 ans et plus),

[3] Depuis le 1er janvier 2011, l'Institut National de Recherche sur les Transports et leur Sécurité (INRETS) et le Laboratoire Central des Ponts et Chaussées (LCPC) ont fusionné pour donner naissance à l'Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux (IFSTTAR).

- **Appartenance de l'immeuble à un organisme HLM** (oui ; non),

- **Vague de l'enquête** (30 avril – 24 juin 2007 ; 25 juin – 2 septembre 2007 ; 3 septembre – 28 octobre 2007 ; 29 octobre – 23 décembre 2007 ; 2 janvier– 2 mars 2008 ; 3 mars – 27 avril 2008).

TABLE 2.1 – *Rapport des côtes de la propension de non-réponse selon les variables expliquant le mécanisme de réponse. Tableau tiré de Roux (2012).*

| Variables | | Rapport des côtes | Pr > $\chi^2$ |
|---|---|---|---|
| Constante | | - 1,17 | <,0001 |
| Âge | De 15 à 34 ans | 0,99 | 0,0362 |
| | De 35 à 49 ans | 0,89 | 0,1479 |
| | De 50 à 64 ans | 0,83 | 0,0003 |
| | 65 ans et plus | 1,00 | Réf. |
| Logement HLM | Non | 0,89 | 0,0186 |
| | Oui | 1,00 | Réf. |
| Type de bâtiment | Logements collectifs | 1,42 | <,0001 |
| | Maison | 1,00 | Réf. |
| Nombre de pièces | 0-1 pièce | 1,60 | <,0001 |
| | 2-3 pièces | 1,24 | 0,3038 |
| | 4-5 pièces | 1,05 | <,0001 |
| | 6 pièces et plus | 1,00 | Réf. |
| Zone de résidence | Commune rurale | 0,64 | <,0001 |
| | Unité urbaine de moins de 20000 habitants | 0,82 | 0,5577 |
| | Unité urbaine de 20000 à 99999 habitants | 0,83 | 0,5300 |
| | Unité urbaine de plus de 100000 habitants | 0,76 | 0,1334 |
| | Unité urbaine de Paris | 1,00 | Réf. |
| Motorisation | Ménage non motorisé | 1,28 | 0,0003 |
| | Ménage avec une voiture | 1,15 | 0,6573 |
| | Ménage avec deux voitures et plus | 1,00 | Réf. |
| Vague | 30 avril – 24 juin 2007 | 0,89 | 0,0167 |
| | 25 juin – 2 septembre 2007 | 1,00 | Réf. |
| | 3 septembre – 28 octobre 2007 | 1,18 | <,0001 |
| | 29 octobre – 23 décembre 2007 | 0,95 | 0,5048 |
| | 2 janvier – 2 mars 2008 | 1,04 | 0,0731 |
| | 3 mars – 27 avril 2008 | 0,83 | <,0001 |

Source : ENTD 2007-2008.

Dans le Tableau 2.1, Roux & Armoogum (2008, 2010, 2011) et Roux (2012) ont ainsi opposé en première analyse :

- *"les ménages habitant une maison aux ménages résidant dans une habitation collective. Les échecs sont plus fréquents pour les logements collectifs (c'est probablement une question d'accessibilité du logement) ;*

- *les ménages habitant un studio ou une chambre aux ménages résidant dans des logements ayant plusieurs pièces. Cette variable est corrélée avec le nombre de personnes vivant dans le ménage. Ainsi, une taille de ménage plus importante s'accompagne d'une probabilité plus grande de réaliser l'entretien ;*

- *les logements situés en zone rurale et en agglomération de moins de 20000 habitants à ceux situés dans l'agglomération de Paris. Les échecs sont d'autant plus nombreux qu'on progresse vers une plus grande urbanisation ;*

- *les ménages n'ayant aucune automobile aux ménages motorissés. Les ménages non-équipés en automobile sont moins favorables à la réalisation des entretiens ;*

- *les ménages dont la personne de référence a moins de 35 ans ou plus de 65 ans à ceux dont l'âge se situe entre 35 et 65 ans. Certainement pour des raisons différentes, les taux d'échec sont plus importants pour les ménages dont la personne de référence a moins de 35 ans et pour ceux dont l'âge de la personne de référence est supérieur à 65 ans. Pour les premiers cela souligne la difficulté des enquêteurs de joindre ces ménages et pour les seconds la réticence des personnes âgées à répondre à un long questionnaire ;*

- *les ménages résidant dans une HLM aux autres. Les échecs sont plus nombreux pour les ménages habitant une HLM ;*

- *les ménages interrogés au mois de juillet–août de ceux interrogés à un autre moment de l'année. Les échecs sont plus nombreux pendant les vacances d'été, période au cours de laquelle nous supposons que les ménages sont les plus mobiles."*

Toujours selon les mêmes auteurs, "*les variables disponibles dans le recensement de 1999 expliquant le mécanisme de réponse de l'*ENTD*, soulignent souvent la difficulté de l'enquêteur à accéder aux logements, même si le digicode n'est semble-t-il pas un facteur aggravant le taux de réponse.*"

Le redressement de l'ENTD a été réalisé en calant l'échantillon des répondants sur les marges disponibles des variables qui expliquent le mécanisme de non-réponse, d'autres variables de calage ont été par ailleurs introduites dans cette étape. Les variables de calage ont été les suivantes :

- **Catégorie socio-professionnelle** de la personne de référence (agriculteur ; indépendant, profession libérale ; professeur, instituteur ... actif ; professeur, instituteur ... à la retraite ; employé, ouvrier ... actif ; employé, ouvrier ... à la retraite ; inactif, chômeur n'ayant jamais travaillé),

- **Sexe** $\times$ **âge de la personne de référence** (homme ; femme) $\times$ (15-29 ans ; 30-39 ans ; 40-49 ans ; 50-59 ans ; 60-69 ans ; 75 ans et plus),

- **Type du ménage** (personne seule ; couple sans enfants ; famille monoparentale ; couple avec enfants ; autres),

- **Type du logement** (maison individuelle ; autres),

- **Nationalité de la personne de référence** (français ; autres),
- **Zone de résidence** des ménages (Paris ; banlieue parisienne ; ville centre de 200000 à 2 millions d'habitants ; banlieue de 200000 à 2 millions d'habitants ; ville centre de 100000 à 200000 habitants ; banlieue de 100000 à 200000 habitants ; ville centre de 50000 à 100000 habitants ; banlieue de 50000 à 100000 habitants ; ville centre de 20000 à 50000 habitants ; banlieue de 20000 à 50000 habitants ; ville centre de moins de 20000 habitants ; banlieue de moins de 20000 habitants ; rurale dans un espace non à dominante rurale ; rurale dans espace à dominante rurale),
- **Nombre d'individus** (nombre d'hommes ; nombre de femmes) $\times$ (0-24 ans ; 25-34 ans ; 35-49 ans ; 50-64 ans ; 65 ans et plus),
- **Motorisation du ménage** (0 voiture, 1 voiture ; 2 voitures et plus),
- **Vague de l'enquête** (30 avril – 24 juin 2007 ; 25 juin – 2 septembre 2007 ; 3 septembre – 28 octobre 2007 ; 29 octobre – 23 décembre 2007 ; 2 janvier – 2 mars 2008 ; 3 mars – 27 avril 2008).

Toutes les procédures de calage ont été réalisées à l'aide de la macro CALMAR2 développée par Le Guennec & Sautory (2002). Des estimations régionales peuvent être ainsi obtenues à partir de ce redressement national. Le Tableau 2.2 donne un exemple des estimations du nombre total de voitures, du nombre de voitures fonctionnant au diesel et du nombre de voitures fonctionnant à l'essence ou autres sources d'énergie, par ménage, au niveau de la région Rhône-Alpes. Ces estimations sont issues du redressement au niveau national de l'ENTD 2007-2008. Pour le calcul des variances, le plan de sondage de l'échantillon national est approché par un plan de Poisson. Ceci est dû au problème d'*"accès à tous les paramètres du tirage de l'échantillon national"*, comme le souligne Le Guennec (2012). De même, le nombre de ménages en Rhône-Alpes est supposé connu.

TABLE 2.2 – *Estimations du nombre total de voitures, du nombre de voitures fonctionnant au diesel et du nombre de voitures fonctionnant à l'essence, par ménage, au niveau de la région Rhône-Alpes, estimations des variances ainsi que des intervalles de confiance à 95% associés, issus du redressement au niveau national de l'*ENTD *2007-2008.*

| **Nombre de voitures** | Estimateur | | Intervalle de confiance à 95% | |
|---|---|---|---|---|
| | Variance | Moyenne | Borne inférieure | Borne supérieure |
| **Total** | 0.00207 | 1.32 | 1.23 | 1.41 |
| **Diesel** | 0.00097 | 0.70 | 0.64 | 0.76 |
| **Essence et autres** | 0.00080 | 0.62 | 0.57 | 0.68 |

Source : ENTD 2007-2008.

Dans la pratique, lorsque la taille de l'échantillon est assez grande, il est

"facile" de satisfaire aux équations de calage. Afin de mener une discussion sur les variables à choisir pour le calage, nous nous sommes restreints à un sous-échantillon, la région Rhône-Alpes. De plus, nous traiterons le cas de plusieurs variables d'intérêts.

### 2.3.3 Choix de partir sur le redressement d'un sous-échantillon

L'échantillon des ménages répondants de l'ENTD 2007-2008 compte 20178 ménages sur toute la France et 986 ménages au niveau de la région Rhône-Alpes. On souhaite estimer, par ménage, le nombre total de voitures particulières, le nombre de voitures utilisant du diesel, et de voitures utilisant de l'essence, au niveau de cette région ainsi que les variances associées. Le Tableau 2.3 donne les matrices de corrélations entre les trois différentes variables d'intérêt considérées, au niveau national et au niveau de la région Rhône-Alpes.

TABLE 2.3 – *Matrices des corrélations entre les trois variables d'intérêt considérées, au niveau national et au niveau de la région Rhône-Alpes.*

| **Nombre de voitures** | | **Nombre de voitures** | | |
|---|---|---|---|---|
| | | **Total** | **Diesel** | **Essence** |
| **Total** | *National* | 1 | 0.82 | 0.76 |
| | *Rhône-Alpes* | 1 | 0.67 | 0.53 |
| **Diesel** | *National* | 0.82 | 1 | 0.24 |
| | *Rhône-Alpes* | 0.67 | 1 | -0.27 |
| **Essence** | *National* | 0.76 | 0.24 | 1 |
| | *Rhône-Alpes* | 0.53 | -0.27 | 1 |

Source : ENTD 2007-2008.

A partir du recensement de la population de 2008, nous disposons de marges connues au niveau de la région Rhône-Alpes. Nous souhaitons savoir quelles sont les variables auxiliaires qui peuvent améliorer nos estimateurs. Les marges disponibles sont : la **motorisation** (ménage sans voiture ; ménage ayant au moins une voiture), le **type du ménage**, l'**âge** de la personne de référence du ménage, le **sexe** de la personne de référence du ménage, la **zone de résidence**, le **type d'aire urbaine de résidence**, le **type du logement**, la **taille du ménage**, la **catégorie socio-professionnelle** de la personne de référence du ménage et la **vague de l'enquête**.

Réaliser un calage avec toutes ces informations auxiliaires amènerait au phénomène de sur-calage. Nous nous posons la question d'un choix judicieux des variables les plus utiles pour améliorer la précision de nos estimateurs.

## 2.4 Choix des variables

Dix variables auxiliaires sont disponibles pour le redressement du sous-échantillon de la région Rhône-Alpes. La procédure de sélection des variables se fait en deux étapes. La première consiste à éliminer les variables auxiliaires non significatives, en ayant recours au critère de l'AIC[4]. En principe, après cette première étape, la variance de l'estimateur calé sur les variables retenues comme significatives devrait être minimale. Cependant, les poids de calage obtenus, utilisés dans le calcul de variance, sont très instables. Les poids initiaux de calage sont très dispersés et peuvent prendre des valeurs très élevées. De plus, le calage est effectué avec des variables catégorielles. Les poids finaux de calage sont en conséquence très dispersés à leur tour et peuvent également prendre des valeurs très élevées malgré l'utilisation de la pseudo-distance de type logistique. En enlevant d'autres variables auxiliaires dans la procédure de calage, les variances peuvent donc encore diminuer. La deuxième étape de la procédure de sélection des variables intervient dans ce cadre. En notant $p$ le nombre de variables auxiliaires retenues comme significatives par le critère de l'AIC, la deuxième étape de la procédure consiste à calculer $p$ variances en n'utilisant dans les calages que $p-1$ variables sur les $p$ à chaque fois, chacune des $p$ variables étant mise de côté une seule fois. Ainsi, les $p-1$ variables associées à la plus petite variance sont retenues si cette nouvelle variance est inférieure à celle obtenue avec les $p$ variables. La procédure est ensuite répétée en calculant $p-1$ variances en n'utilisant dans les calages que $p-2$ variables sur les $p-1$ à chaque fois, chacune des $p-1$ variables étant mise de côté une seule fois. Les $p-2$ variables associées à la plus petite variance sont retenues si cette nouvelle variance est inférieure à celle obtenue avec les $p-1$ variables. Et ainsi de suite. Dans le cas où la nouvelle variance n'est pas inférieure à celle obtenue précédemment, deux variables sont simultanément mises de côté, puis si nécessaire trois variables simultanément, ..., jusqu'à $p-1$ variables simultanément si nécessaire.

L'Algorithme 2.1 est proposée afin de sélectionner les variables auxiliaires pertinentes pour le redressement du sous-échantillon de la région Rhône-Alpes.

Pour chacune des variables d'intérêt considérées, le Tableau 2.4 résume la comparaison entre les variances minimales obtenues par la procédure de sélection et les variances obtenues par calage global sur toutes les variables auxiliaires disponibles au niveau de la région Rhône-Alpes ainsi que les variances obtenues avec le redressement national. Le Tableau 2.5 présente

[4]Le critère d'Akaike (AIC) est défini par la formule : $AIC = 2k - 2\ln\mathcal{L}$ où $k$ est le nombre de paramètres dans le modèle considéré et $\mathcal{L}$ est la fonction de vraisemblance.

un récapitulatif des différentes variables de calage utilisées pour obtenir les variances optimales pour chacune des variables d'intérêt considérées.

**Algorithme 2.1:** Procédure de sélection des variables pertinentes

1 Retenir les 10 variables de calage et calculer la variance de l'estimateur obtenu par calage sur les 10 variables;
2 Par le critère de l'AIC, déterminer les variables considérées comme non significatives;
3 Retenir les variables de calage significatives restantes;
4 **tant que** *La variance diminue* **faire**
5 Retirer à chaque fois une variable et calculer la variance associée à l'estimateur obtenu;
6 Retenir les variables de calage restantes donnant la plus petite variance et tant qu'elle reste inférieure à la variance du précédent modèle
7 **pour** *i allant de 2 à nombre de variables de calage restantes* **faire**
8 **si** *La variance ne diminue pas* **alors**
9 refaire 5 - 6 avec *i* variables simultanément

TABLE 2.4 – *Comparaison entre les variances obtenues avec le redressement national, les variances minimales obtenues par la procédure de sélection et les variances obtenues par calage global sur toutes les variables auxiliaires disponibles au niveau de la région Rhône-Alpes.*

| **Nombre de voitures** | Estimation de la variance | | | Gain de précision entre la variance par calage global et la variance minimale (en %) |
|---|---|---|---|---|
| | Redressement national | Redressement régional | | |
| | | calage global Rhône-Alpes | variance minimale | |
| **Total** | 0.00207 | 0.00053 | 0.00042 | 20.75 |
| **Diesel** | 0.00097 | 0.00065 | 0.00056 | 13.85 |
| **Essence** | 0.00080 | 0.00055 | 0.00046 | 16.36 |

Source : ENTD 2007-2008.

### 2.4.1 Le nombre total de voitures

La variance de l'estimateur obtenu par le calage global sur les 10 variables est de 0.00053. Dans un premier temps, quatre variables de calage sont considérées comme non-significatives pour expliquer la variable d'intérêt **nombre total de voitures**, par le critère de l'AIC. Classées par ordre de non-significativité, celles-ci sont les variables **sexe**, **vague**, **type de l'aire urbaine**, **catégorie socio-professionnelle**. Par la procédure de sélection proposée, les variables retenues sont : **motorisation**, **type du ménage**, **âge**, **type du logement**, **taille du ménage**. La variance de l'estimateur obtenu par le calage avec ces variables retenues est de 0.00042.

### 2.4.2 Le nombre de voitures diesel

La variance de l'estimateur obtenu par le calage sur les 10 variables de calage est de 0.00065. Par le critère de l'AIC, seule variable zone de résidence n'est pas significative pour expliquer la variable d'intérêt **nombre de voitures diesel**. Les variables retenues par la méthode de sélection proposée sont : **motorisation**, **type du ménage**, **âge**, **sexe**, **type du logement**, **taille du ménage**, **catégorie socio-professionnelle** et **vague**. La variance de l'estimateur obtenu par le calage avec ces variables est de 0.00056.

### 2.4.3 Le nombre de voitures essence

La variance de l'estimateur obtenu par le calage global sur les 10 variables est de 0.00055. Dans un premier temps, quatre variables de calage sont considérées comme non-significatives pour expliquer la variable d'intérêt **nombre de voitures essence**, par le critère de l'AIC. Classées par ordre de non-significativité, celles-ci sont les variables **type du ménage**, **sexe**, **type de l'aire urbaine**. Par la procédure de sélection proposée, les variables retenues sont : **motorisation** et **type du logement**. La variance de l'estimateur obtenu par le calage avec ces variables retenues est de 0.00046.

Table 2.5 – *Variables auxiliaires sélectionnées pour l'optimisation des variances.*

| | Nombre de voitures | | |
|---|---|---|---|
| | **Total** | **Diesel** | **Essence et autres** |
| **Motorisation** | x | x | x |
| **Type du ménage** | x | x | |
| **Sexe** | | x | |
| **Âge** | x | x | |
| **Zone de résidence** | | | |
| **Type de l'aire urbaine** | | | |
| **Type du logement** | x | x | x |
| **Taille du ménage** | x | x | |
| **Catégorie socio-professionnelle** | | x | |
| **Vague** | | x | |

Source : Entd 2007-2008.

## 2.5 Discussion

Par la procédure de sélection de variables, les variances optimales sont obtenues avec des jeux différents de variables auxiliaires pour chacune des variables d'intérêt considérées. Par commodité, il serait judicieux de ne considérer qu'un unique système de pondération, quitte à diminuer sensiblement la précision des estimations obtenues. L'utilisation du système de pondération issu du calage avec les variables auxiliaires sélectionnées pour l'optimisation de la variance pour la variable **nombre**

**total de voitures** semble être un bon compromis. Les variables retenues pour un unique système de pondération sont donc : **motorisation**, **type du ménage**, **âge**, **type du logement**, **taille du ménage**.

TABLE 2.6 – *Variances obtenues par calage global sur toutes les variables auxiliaires disponibles, variances minimales obtenues par la procédure de sélection, variances obtenues en utilisant le même système de pondération pour les trois variables d'intérêt considérées, et gains de précision (en %).*

| **Nombre de voitures** | Estimation de la variance | | | Gain de précision | |
|---|---|---|---|---|---|
| | calage global (1) | variance minimale (2) | après compromis (3) | entre (1) et (2) | entre (2) et (3) |
| **Total** | 0.00053 | 0.00042 | 0.00042 | 20.75 | 0 |
| **Diesel** | 0.00065 | 0.00056 | 0.00057 | 12.31 | -1.79 |
| **Essence** | 0.00055 | 0.00046 | 0.00049 | 10.91 | -6.52 |

Source : ENTD 2007-2008.

Le Tableau 2.6 donne une comparaison des différentes variances obtenues par calage global sur toutes les variables auxiliaires disponibles au niveau de la région Rhône-Alpes, par la procédure d'optimisation, par l'utilisation d'une même pondération, ainsi que les gains de précision après l'utilisation de ce même système de pondération. La précision de la variable **nombre de voitures diesel** ne diminue pas très sensiblement après l'utilisation de l'unique système de pondération, le gain de précision n'allant que de 13.85% à 12.31% et la variance n'augmentant que d'un cent-millième. La précision de la variable **nombre de voitures essence** diminue plus sensiblement avec un gain de précision allant de 16.36% à 10.91% et avec une variance augmentant de trois cent-millièmes.

La variable **zone de résidence** n'est en aucun cas sélectionnée durant la procédure d'optimisation pour les trois variables d'intérêt considérées. Il est intéressant de voir comment se comportent les variances lorsque l'on croise la variable **zone de résidence** avec les variables d'intérêt. Le Tableau 2.7 montre un exemple de la précision des estimations, issues du système unique final obtenu de pondérations, pour les ménages résidant dans les zones à dominante rurale et dans les zones à dominante urbaine. Le Tableau 2.8 montre le même exemple issu d'un calage uniquement sur la variable **zone de résidence**. Au niveau des zones de résidence à dominante urbaine, les estimations sont meilleures en utilisant le système de pondération unique obtenu après la sélection des variables. Au niveau des zones de résidence à dominante rurale, les précisions des estimations sont sensiblement équivalentes.

TABLE 2.7 – *Variances des estimations issues d'un même système de pondération au niveau des ménages résidant dans les zones à dominante rurale et à dominante urbaine.*

| **Nombre de voitures** | Estimation de la variance au niveau des zones de résidence | |
|---|---|---|
| | à dominante rurale | à dominante urbaine |
| **Total** | 0.00027 | 0.00062 |
| **Diesel** | 0.00012 | 0.00060 |
| **Essence** | 0.00009 | 0.00049 |

Source : ENTD 2007-2008.

TABLE 2.8 – *Variances des estimations issues d'un calage sur la variable **zone de résidence** au niveau des ménages résidant dans les zones à dominante rurale et à dominante urbaine.*

| **Nombre de voitures** | Estimation de la variance au niveau des zones de résidence | |
|---|---|---|
| | à dominante rurale | à dominante urbaine |
| **Total** | 0.00017 | 0.00087 |
| **Diesel** | 0.00012 | 0.00065 |
| **Essence** | 0.00007 | 0.00053 |

Source : ENTD 2007-2008.

## 2.6 Conclusion

La méthode présentée dans ce papier dépend des variables auxiliaires disponibles ainsi que de leur pouvoir explicatif sur les variables d'intérêt considérées. Dans le cas du sous-échantillon de la région Rhône-Alpes de l'ENTD 2007-2008, les variances optimales obtenues pour le nombre total de voitures particulières, le nombre de voitures utilisant du diesel, et de voitures utilisant de l'essence, dépendent clairement du choix des variables auxiliaires utilisées dans les procédures de calage. Afin d'éviter différentes pondérations distinctes pour chacune de ces variables d'intérêt, un système de pondération unique résultant d'un calage sur les variables **motorisation**, **type du ménage**, **âge**, **type du logement**, **taille du ménage** a été établi. Les précisions finales obtenues avec ce système unique obtenu ont la caractéristique d'être assez équivalentes avec les précisions minimales résultant de la procédure de sélection des variables auxiliaires pertinentes pour chacune des variables d'intérêt.

L'inconvénient de la méthode présentée est que la précision des estimations est faible dès lors que la taille du sous-échantillon, inclus dans un domaine, est petite. Dans ce cas, il est donc nécessaire d'emprunter de la "force" en dehors du domaine considéré.

# 3 Small Area Estimation by Splitting the Sampling Weights

**Abstract** A new method is proposed for small area estimation. The principle is based upon the splitting of the sampling weights between the areas. A matrix of weights is defined. Each column of this matrix enables us to estimate the totals of the variables of interest at the level of an area. This method automatically satisfies the coherence property between the local estimates and the overall estimate. Moreover, the local estimators are calibrated on auxiliary information available at the level of the small areas. This methodology also enables the use of composite estimators that are weighted means between a direct estimator and a synthetic estimator. Once the weights are computed, the estimates can be easily computed for any variable of interest. A set of simulations shows the interest of the proposed method.[1]

**Keywords** indirect estimator, matrix calibration, weights

## 3.1 Introduction

Three main families of estimators are usually used by statisticians to increase the quality of estimates at the level of small areas: direct estimators and indirect estimators based on implicit or explicit models (see, for instance, Rao, 2003).

The direct estimators are based on the survey data provided only by the considered area. When available, the auxiliary information only depends on the units of the area. The family of direct estimators gathers

[1]This chapter is a reprint of: Randrianasolo & Tillé (2013). Small area estimation by splitting the sampling weights. *Electronic Journal of Statistics* 7 1835–1855. http://dx.doi.org/10.1214/13-EJS827.

the estimator proposed by Horvitz & Thompson (1952) also called $\pi$-estimator, the generalized regression estimators (Särndal et al., 1992) and the calibration estimators (Deville & Särndal, 1992). The main problem with this family of estimators is the increasing variance when the area size decreases. A wise choice of auxiliary information can reduce the variance.

The indirect estimators can depend on all the sampled units for the estimation of a particular area. These estimators are based on the notion of deriving strength from space because a unit of a given area can help for the estimation of any area. Two sub-families of estimators are distinguished: synthetic and composite estimators.

According to Gonzalez (1973), *"an unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for sub-areas on the assumption that the small areas have the same characteristics as the larger area, we identify these estimates as synthetic estimates"*. Practically, these estimators are based on the hypothesis of the equality of a parameter within the areas and in the overall population. The synthetic estimator can thus be a prediction by means of a linear model of a total based upon the assumption that the regression coefficient does not vary from an area to another. These estimators generally have a low variance since they depend on all the observations, i.e. the size of the sample is thus large. Unfortunately, these estimators can miss a specificity of a given area. The composite estimators are weighted means of a direct estimator and a synthetic one. For each area, the weight of the mean can be chosen by minimizing the mean squared error.

The indirect estimators built from explicit modeling are based on linear mixed models, generalized linear mixed models and Bayesian procedures which derive Best Linear Unbiased predictors (BLUP), Empirical Best Linear Unbiased Predictors (EBLUP) and empirical Bayes estimators (see, among others, Fuller & Battese, 1973; Prasad & Rao, 1990, 1999; Rao, 2003). The most famous model using linear mixed models is the one developed by Fay & Herriot (1979). The authors begin by modeling a function of the mean in a given area, explained on the one hand, by the auxiliary information, and on the other hand, by a random part explaining the variability across areas that are not considered in the auxiliary information. Then, they show that the BLUP is a composite estimator.

In small area estimation, an important factor to bear in mind is that local estimates are not always consistent with the overall population estimate. Indeed, in general, the sum of estimates at the level of small areas does not coincide with the estimate at the level of the overall population. In order to satisfy a benchmarking property, Prasad & Rao (1999) propose

a two-step procedure to obtain a pseudo-EBLUP of a small area mean. They combine area models using survey weights with unit level models. Using a nested error regression model, You & Rao (2002) developed a method that provides coherent estimates thanks to a skillful variable change in the regression and the use of sampling weights so as to build a pseudo-EBLUP. Also, under a nested error regression model, You & Rao (2003) use a pseudo-hierarchical Bayes approach to obtain posterior estimators of small area means. Ugarte et al. (2009) propose an EBLUP based upon a linear mixed model with restrictions. They force the sum of small area estimates to equal the calculated estimate of the overall population using a synthetic estimator.

In this paper, a new approach is proposed. The method consists of splitting the sampling weights of the overall population estimator to construct local estimators. Each weighting system corresponds to a small area. The idea is to define weights for the areas that depend on the sampled units as well as the other small areas. Each statistical unit can contribute to all the small areas.

In order to satisfy a benchmarking principle, the sum of the weights of a particular unit relative to each area must be equal to its global weight, which automatically implies that the global estimator is the sum of the local estimates. Furthermore, the weights are calibrated in such a way that, in each area, the estimates of the totals are equal to the population total for the auxiliary variables that are known at the level of the small areas.

The main tool of this method is a matrix $\mathbf{Q}$ for which the number of rows is equal to the number of units and the number of columns is equal to the number of small areas. This matrix embodies the way the global weights are split into the areas. The benchmarking principle is obtained by the simple fact that the sum of the elements of each row is equal to 1. For each unit, two kinds of contributions are distinguished: its contribution to its own area ("auto-contribution") and its contribution to the other areas ("extra-contribution"). A consequence of the benchmarking principle is that the more a unit contributes to its own area, the less it contributes to the other ones. A composite estimator is built with an "area" part by a direct estimation and with an "extra-contribution" part, given the $\mathbf{Q}$ probability matrix. At the level of each small area, these two parts are balanced with a parameter (tuning constant) obtained by minimizing the statistical dispersion of the variable of interest.

This paper is structured as follows. In Section 3.2, the notation is defined. The direct estimators are presented in Section 3.3. In Section 3.4, the weights splitting estimation is developed. Next, in Section 3.5, the

composite estimator is presented. The choice of the tuning constant is discussed in Section 3.6. A simulation study is presented in Section 3.7, and the paper ends with some brief concluding remarks in Section 3.8.

## 3.2 Notation

Consider a finite population $U$ of $N$ statistical units belonging to $D$ disjoint areas $\{A_1, \ldots, A_d, \ldots, A_D\}$ of sizes $\{N_1, \ldots, N_d, \ldots, N_D\}$. The units can be identified by a label $k \in \{1, \ldots, k, \ldots, N\}$. Consider also $J$ auxiliary variables $x_1, \ldots, x_j, \ldots, x_J$. We are interested in estimating the overall total

$$t_y = \sum_{k \in U} y_k.$$

of the interest variable $y$. Moreover, we want to estimate the total of $y$ in each area, i.e. for area $A_d$

$$t_y^d = \sum_{k \in U \cap A_d} y_k.$$

The estimation is based upon the availability of $J$ auxiliary variables. The values of the $j$th variable for all the units are denoted $x_{1j}, \ldots, x_{kj}, \ldots, x_{Nj}$. The values of the $J$ variables of unit $k$ are denoted by the column vector of $\mathbb{R}^J$

$$\mathbf{x}_k = \begin{pmatrix} x_{k1} & \cdots & x_{kj} & \cdots & x_{kJ} \end{pmatrix}^\top.$$

The $J$ variables of the population are represented by $N \times J$ matrix

$$\mathbf{X}_U = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_k & \cdots & \mathbf{x}_N \end{pmatrix}^\top.$$

The totals of the $J$ variables for the population are represented by the row vector of size $J$

$$\mathbf{t}_x = \begin{pmatrix} \sum_{k \in U} x_{k1} & \cdots & \sum_{k \in U} x_{kj} & \cdots & \sum_{k \in U} x_{kJ} \end{pmatrix}.$$

The totals of the $J$ variables for the population belonging to area $A_d$ are represented by the row vector of size $J$

$$\mathbf{t}_x^d = \begin{pmatrix} \sum_{k \in U \cap A_d} x_{k1} & \cdots & \sum_{k \in U \cap A_d} x_{kj} & \cdots & \sum_{k \in U \cap A_d} x_{kJ} \end{pmatrix},$$

and the totals of the $J$ variables for the population belonging to each of the areas are represented by the matrix of size $D \times J$

$$\mathbf{t}_x^A = \begin{pmatrix} \mathbf{t}_x^1 \\ \vdots \\ \mathbf{t}_x^d \\ \vdots \\ \mathbf{t}_x^D \end{pmatrix}. \tag{3.1}$$

Below, matrix $\mathbf{t}_x^A$ is supposed to be known. This matrix can thus be used to improve estimation of the totals in domains $t_y^d$.

A sample $s$ is a subset of $U$, and a sampling design $p(s)$ is a probability distribution on all possible samples that can be drawn from $U$, such that

$$p(s) \geq 0, \text{ and } \sum_{s \subset U} p(s) = 1.$$

For a given sampling design $p(s)$, a sample $s$ is the realization of a random sample $S$, i.e. $\Pr(S = s) = p(s)$ for all $s \subset U$. We note $\{n_1, \ldots, n_d, \ldots, n_D\}$ the sizes of the areas $\{\#(S \cap A_1), \ldots, \#(S \cap A_d), \ldots, \#(S \cap A_D)\}$. The first order inclusion probability of unit $k$ is denoted by $\pi_k = \Pr(k \in S)$.

## 3.3 Direct estimation

At the level of an area, direct estimation consists of building an estimator of $t_y^d$ without using any information outside of the given area. Then, in a direct estimation, a unit can only contribute to its own area. For instance, the Horvitz & Thompson (1952) estimator (or $\pi$-estimator) which directly uses the sample weights $1/\pi_k$ is a direct estimator. A calibrated estimator is also a direct estimator.

Let $t_y = \sum_{k \in U} y_k$ be the total of the quantitative variable $y$. The "Horvitz & Thompson (1952) estimator" or "$\pi$-estimator" of $t_y$ is defined by

$$\hat{t}_{y,\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

At the level of area $A_d$, the Horvitz & Thompson (1952) estimator of the total of the variable $y$ denoted by $t_y^d = \sum_{k \in U \cap A_d} y_k$ is given by the quantity:

$$\hat{t}_{y,\pi}^d = \sum_{k \in S \cap A_d} \frac{y_k}{\pi_k}.$$

This method of calibration was formalized by Deville & Särndal (1992) who gave a common framework for the calibration estimation and the properties of these estimators. Suppose that a vector of totals $\mathbf{t}_x$ of $J$ auxiliary variables is known at the level of the population. The calibration estimator of the total of the quantitative variable $y$ depends on a weighting system $w_k$. The estimator is defined by

$$\hat{t}_{y,w} = \sum_{k \in S} w_k y_k,$$

where the $w_k$ weighting system depends on the sample $S$ and satisfies the calibration equation:

$$\sum_{k \in S} w_k \mathbf{x}_k^\top = \mathbf{t}_x. \tag{3.2}$$

The calibration equation (3.2) means that the weighting system of the calibrated estimator must reproduce exactly the values of the totals of the auxiliary variables that are known at the population level.

The weights $w_k$ are computed in such a way to be as close as possible to the Horvitz-Thompson weights $d_k = 1/\pi_k$. Moreover, the weights must satisfy the calibration equation (3.2). In order to find such weights, Deville & Särndal (1992) propose several pseudo-distances denoted by $G_k(w_k, d_k)$ that is assumed to be positive, derivable, strictly convex with regard to $w_k$ and such that $G_k(d_k, d_k) = 0$ for all $k \in U$. The $w_k$ weights are obtained by minimizing the quantity

$$\sum_{k \in S} \frac{G_k(w_k, d_k)}{q_k}$$

subject to the constraints of the calibration equation given in (3.2). The ${q_k}^{-1}$ are coefficients which determine the importance of each unit in the calculation of the distance.

Many different distances can be used and are discussed in Deville & Särndal (1992). In general, if $g_k(w_k, d_k)$ denotes the derivative of $G_k(w_k, d_k)$ with respect to $w_k$, then the weights are defined by

$$w_k = d_k F_k(q_k \boldsymbol{\lambda}^\top \mathbf{x}_k), \tag{3.3}$$

where $\boldsymbol{\lambda}$ is the vector of Lagrangian multiplier and $d_k F_k(.)$ is the reciprocal function of $g_k(., d_k)$. The value of the Lagrangian multipliers $\boldsymbol{\lambda}$ can be identified by inserting (3.3) in (3.2) and by solving the calibration equation by the Newton-Raphson method.

Below, we mainly use the raking-ratio method which is defined by means of the Kullback-Leibler measure. In this case, $q_k = 1, k \in S$

$$G_k(w_k, d_k) = w_k \log \frac{w_k}{d_k} + d_k - w_k,$$

$$g_k(w_k, d_k) = \log \frac{w_k}{d_k},$$

and

$$w_k = d_k F_k(q_k \boldsymbol{\lambda}^\top \mathbf{x}_k) = d_k \exp(\boldsymbol{\lambda}^\top \mathbf{x}_k).$$

At the level of small areas, a direct estimator cannot be used when the sample size within the area is small because its variance becomes very large. For a given area $A_d$ of size $n_d$, the variance of a direct estimator is $\mathcal{O}(1/n_d)$. The smaller the size $n_d$, the larger the variance. Hence, the quality of small area direct estimates is debatable. When the size of a given area is not large enough to have a satisfactory direct estimation, we attempt to improve the quality of the estimates by borrowing information at the level of the other areas. And so we use the method of splitting the sampling weights.

## 3.4 Weights splitting or extra-contribution estimation

### 3.4.1 Constraints on the split weights

The proposed method consists of splitting the weights into the areas. In a direct estimator, only the weights of the units that belong to an area $A_d$ can contribute to a local estimation at the level of $A_d$, which is called "auto-contribution part". In the weights splitting method, any unit can contribute to the estimation of any area through a certain weight, which is called "extra-contribution part".

Suppose that a weighting system has already been computed for the overall estimation. These weights can be the inverse of the inclusion probabilities or can be obtained by means of a calibration procedure on the total $\mathbf{t}_x$. The main idea of the proposed weight-splitting approach is to build a weight $w_{kd}$ which depends both on unit $k$ and area $A_d$. This weight is defined as the product of the basic $w_k$ weight with a splitting coefficient $q_{kd}$ that distributes the weights in the areas, i.e.

$$w_{kd} = w_k q_{kd}, \text{ for all } k \in S \text{ and for } d = 1, \ldots, D.$$

The $D \times J$ matrix $\mathbf{t}_x^A$ of the $D$ totals of the auxiliary variables, given in (3.1) is supposed to be known. The knowledge of this auxiliary information at the level of the areas is used to calibrate the weighting system.

More precisely we would like to have weights for the areas that are calibrated on the totals of the areas, i.e.

$$\sum_{k \in S} w_{kd} \mathbf{x}_k^\top = \sum_{k \in S} w_k q_{kd} \mathbf{x}_k^\top = \mathbf{t}_x^d, \text{ for all } d, = 1, \ldots, D. \tag{3.4}$$

Moreover, we want to impose a coherence between the sum of small areas estimates and the overall estimates. Since $w_{kd} = w_k q_{kd}$, the coherence

$$\sum_{d=1}^{D} \sum_{k \in S} w_{kd} \mathbf{x}_k^\top = \sum_{k \in S} w_k \mathbf{x}_k^\top$$

can be obtained if

$$\sum_{d=1}^{D} w_{kd} = \sum_{d=1}^{D} w_k q_{kd} = w_k, \text{ for all } k \in S.$$

The splitting coefficients must thus satisfy

$$\sum_{d=1}^{D} q_{kd} = 1. \tag{3.5}$$

To sum up, the $n \times D$ weights $q_{kd}$ must satisfy $D \times J + n$ constraints:

- the $D \times J$ constraints of calibration on the totals of the areas given in (3.4),
- the $n$ constraints of consistency given in (3.5).

Note also that the constraints of consistency given in (3.5) are based on an important practical interpretation. If, in small area estimation, the estimate is written in the form of a weighting system, the constraints of consistency imply that units that strongly contribute to other areas contribute less to their own area. Extra-contribution works like an exchange of information, if there are few units in an area, this area needs to borrow strength from other areas, but in exchange, the units of this small area must contribute more to the other areas.

The weights $q_{kd}$ can be gathered in a $\mathbf{Q}$ matrix with $n$ rows and $D$ columns. The sum of the elements in each row is thus equal to 1:

$$\mathbf{Q} = \begin{pmatrix} q_{11} & \cdots & q_{1d} & \cdots & q_{1D} \\ \vdots & & \vdots & & \vdots \\ q_{k1} & \cdots & q_{kd} & \cdots & q_{kD} \\ \vdots & & \vdots & & \vdots \\ q_{n1} & \cdots & q_{nd} & \cdots & q_{nD} \end{pmatrix},$$

which can be written

$$\mathbf{Q}\mathbf{1}_D = \mathbf{1}_n,$$

where $\mathbf{1}_D$ (resp. $\mathbf{1}_n$) is a column vector of $D$ ones (resp. of $n$ ones).

The constraints given in (3.4) can be rewritten with a matrix notation:

$$\begin{pmatrix} q_{11} & \cdots & q_{k1} & \cdots & q_{n1} \\ \vdots & & \vdots & & \vdots \\ q_{1d} & \cdots & q_{kd} & \cdots & q_{nd} \\ \vdots & & \vdots & & \vdots \\ q_{1D} & \cdots & q_{kD} & \cdots & q_{nD} \end{pmatrix} \begin{pmatrix} w_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & w_k & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & w_n \end{pmatrix} \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_k^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} \mathbf{t}_x^1 \\ \vdots \\ \mathbf{t}_x^d \\ \vdots \\ \mathbf{t}_x^D \end{pmatrix}.$$

In short, the two following properties must be satisfied:

1. the sum of the rows of $\mathbf{Q}$ must equal 1,
2. $\mathbf{Q}^\top \operatorname{diag}(w_1, \ldots, w_k, \ldots, w_n)\mathbf{X}_S = \mathbf{t}_x^A$, where $\mathbf{X}_S = (x_{kj})_{k \in S, j=1,\ldots,J}$.

Each coefficient $q_{kd}$ embodies the contribution of unit $k$ to the estimator for area $A_d$.

### 3.4.2 Computation of matrix Q

In order to compute a matrix $\mathbf{Q}$ which satisfies both the coherence property and the totals at the level of each area, we propose a simple algorithm which repeats two successive calibrations. The columns on the totals of the areas are calibrated on the values of the $x$-variables. Next, the rows of the matrix are calibrated again to ensure that the sum is equal to 1. These steps are repeated until convergence. This method is a generalization of the raking-ratio method that enables us to calibrate a contingency table on marginal totals (see for instance Ireland & Kullback, 1968; Arora & Brackstone, 1977).

More specifically, the algorithm begins with the initialization of the matrix by

$$\mathbf{Q}^{\{0\}} = \begin{pmatrix} \frac{N_1}{N} & \cdots & \frac{N_d}{N} & \cdots & \frac{N_D}{N} \\ \vdots & & \vdots & & \vdots \\ \frac{N_1}{N} & \cdots & \frac{N_d}{N} & \cdots & \frac{N_D}{N} \end{pmatrix}.$$

This first matrix of splitting coefficients simply shares the weights proportionally to the size of the areas in the population.

Next, at step $2t$, for $t = 1, 2, 3, \ldots$ the following two operations are repeated:

- Each column of $\mathbf{Q}^{\{2t-2\}}$ is calibrated on the vector of known totals of each area by solving in $\boldsymbol{\lambda}$ for each area $A_d$, $d = 1 \ldots D$ the equations system:

  $$\mathbf{t}_x^d = \sum_{k \in S} w_k q_{kd}^{\{2t-2\}} \mathbf{x}_k^\top \exp\left(\mathbf{x}_k \boldsymbol{\lambda}_d\right)$$

  The coefficients of the new matrix $\mathbf{Q}^{\{2t-1\}}$ can be obtained by:

  $$q_{kd}^{\{2t-1\}} = q_{kd}^{\{2t-2\}} \exp(\mathbf{x}_k \boldsymbol{\lambda}_d).$$

- Then the sum of the rows is calibrated so as to equal 1:

  $$q_{kd}^{\{2t\}} = \frac{q_{kd}^{\{2t-1\}}}{\sum_{d=1}^{D} q_{kd}^{\{2t-1\}}}.$$

The iteration stops when the sum of the rows is almost equal to 1 after a column calibration, or more specifically when

$$\sum_{k=1}^{n} \left| \sum_{d=1}^{D} q_{kd}^{\{2t-1\}} - 1 \right| < \varepsilon,$$

where $\varepsilon$ is a sufficiently small positive real.

The use of an exponential calibration function guarantees that the weights $q_{kd}$ remain nonnegative at each step of the method. Once the $\mathbf{Q}$ matrix is computed, the totals of areas for any variable of interest can be estimated. At the level of a given area $A_d$, the extra-contribution estimator of the total of a quantitative variable $y$ is defined by

$$\hat{t}^d_{y,q} = \sum_{k \in S} w_k q_{kd} y_k. \tag{3.6}$$

The sum of the area estimators is always equal to the estimator in the population.

## 3.5 Composite estimator

The weights splitting method is a synthetic estimator because all the statistical units can contribute to each area. In order to avoid to miss some area specificities, a composite estimator can be obtained by mixing the direct estimator with the extra-contribution estimator so as to propose a method where each area is estimated by a part of auto-contribution built from a direct estimation and by a part of extra-contribution built thanks to the $\mathbf{Q}$ matrix.

A matrix $\mathbf{C}$ of new composite weights is constructed by means of weights $\alpha_d$ for $d = 1, \ldots, D$. The procedure starts with the construction of a $n \times D$ matrix $\mathbf{G} = (g_{kd})$, where

$$g_{kd} = \alpha_d q_{kd} + (1 - \alpha_d) \mathbb{1}_{\{k \in A_d\}}, k \in S, d = 1, \ldots, D,$$

where $\mathbb{1}_{\{C\}}$ equals 1, if condition $C$ is true and 0 otherwise. The coefficients $\alpha_d$ depend on the areas. The smaller the area, the larger the $\alpha_d$. It is desirable that, in large areas, the estimator depends more on the units of these areas. Whereas in small areas, the estimator depends more on the units outside of these areas.

Next, matrix $\mathbf{G}$ is calibrated again on the two sets of constraints in such a way that the totals of the areas are reproduced for the auxiliary variables and that the sum of the rows equals 1. The algorithm described in Section 3.4.2 is thus applied again. We obtain matrix $\mathbf{C} = (c_{kd})$ the elements of which can be written as $c_{kd} = g_{kd} h_{kd}$ where the $h_{kd}$ are the matrix calibration adjustments.

Considering a quantitative variable of interest $y$, the composite estimator of the total of $y$ at the level of $A_d$ is a weighted average given

by:

$$\begin{aligned}\hat{t}^d_{y,c} &= \sum_{k\in S} c_{kd} w_k y_k \\ &= \alpha_d \sum_{k\in S} h_{kd} q_{kd} w_k y_k + (1-\alpha_d) \sum_{k\in S\cap A_d} h_{kd} w_k y_k. \end{aligned} \tag{3.7}$$

The estimator is a weighted average of two terms. The first one is a synthetic estimator that depends on all the statistical units of the sample. The second one is a direct estimator that only depends on the selected units in the small area.

## 3.6 Determination of a tuning constant $\alpha_d$

### 3.6.1 Approximation of the variance of the composite estimator

In order to obtain a reasonable value for the tuning constants, one can use heuristic reasoning. Since the first term of the composite estimator given in (3.7) depends on all the units, we assume that its variance can be written $\sigma^2_{d,1}/n$. Since the second term only depends on the units that belong to $A_d$, we assume that its variance is equal to $\sigma^2_{d,2}/n_d$.

Moreover, if we assume that the correlation coefficient between the first and the second term is equal to $\rho$, the variance of the composite estimator is equal to the following quantity:

$$\mathrm{Var}(\hat{t}^d_{y,c}) = \alpha_d^2 \frac{\sigma^2_{d,1}}{n} + (1-\alpha_d)^2 \frac{\sigma^2_{d,2}}{n_d} + 2\alpha_d(1-\alpha_d)\rho \frac{\sigma_{d,1}\sigma_{d,2}}{\sqrt{n_d n}}.$$

If we assume that the covariance term is negligible,

$$\mathrm{Var}(\hat{t}^d_{y,c}) \approx \alpha_d^2 \frac{\sigma^2_{d,1}}{n} + (1-\alpha_d)^2 \frac{\sigma^2_{d,2}}{n_d}. \tag{3.8}$$

This kind of approximation is done for composite estimators for instance in Rao (2003, p. 57).

By setting the derivative of (3.8) with respect to $\alpha_d$ to zero, we obtain:

$$\alpha_d \frac{\sigma^2_{d,1}}{n} - (1-\alpha_d)\frac{\sigma^2_{d,2}}{n_d} = 0.$$

The value for $\alpha_d$ that minimizes (3.8) is then given by:

$$\alpha_d(n_d) \approx \frac{1}{1 + \frac{\sigma^2_{d,1}}{\sigma^2_{d,2}} \frac{n_d}{n}}.$$

If

$$\theta_d = \frac{\sigma^2_{d,1}}{\sigma^2_{d,2}},$$

we can see that when $n_d$ tends to 0 (resp. $+\infty$), then $\alpha_d(n_d)$ tends to 1 (resp. 0). If $\sigma^2_{d,1}$ and $\sigma^2_{d,2}$ do not depend on $d$, then we obtain a simplification:

$$\alpha_d(n_d) \approx \frac{1}{1+\theta\frac{n_d}{n}}.$$

### 3.6.2 EBLUP and pseudo-EBLUP under a mixed model

#### Nested error linear regression model

Let us now assume that we are interested in small area means. Henderson (1975); Battese et al. (1988); Prasad & Rao (1990); You & Rao (2002); Rao (2003) proposed a nested error linear regression model to estimate small area means. Using the Prasad & Rao (1990); You & Rao (2002); Rao (2003) notation, we can consider the mixed model approach

$$y_{dk} = \mathbf{x}_{dk}^\top \boldsymbol{\beta} + v_d + \varepsilon_{dk} \tag{3.9}$$

where $k = 1, \ldots, n_d$, $d = 1, \ldots, D$, $v_d$ are independent centered normal variables with variances $\sigma_v^2$, $\varepsilon_{dk}$ are independent centered normal variables with variances $\sigma_\varepsilon^2$. Moreover, the $v_d$ are assumed to be independent from the $\varepsilon_{dk}$.

The mean for area $A_d$, denoted $\overline{Y}_d$, can be approximated by the parameter

$$\mu_d = \overline{\mathbf{X}}_d^\top \boldsymbol{\beta} + v_d$$

where

$$\overline{\mathbf{X}}_d = \frac{1}{N_d}\sum_{k=1}^{N_d} \mathbf{x}_{dk}.$$

You & Rao (2002) proposed a combination of the basic unit level model (3.9) with sample weights and obtained the following weighted area level model

$$\overline{y}_{dw} = \sum_{k=1}^{n_d} \frac{y_{dk}}{\pi_{dk}\sum_{l=1}^{n_d}\frac{1}{\pi_{dl}}} = \overline{\mathbf{x}}_{dw}^\top \boldsymbol{\beta} + v_d + \overline{\varepsilon}_{dw} \tag{3.10}$$

where

$$\mathbb{E}(\overline{\varepsilon}_{dw}) = 0$$

and

$$\mathrm{Var}(\overline{\varepsilon}_{dw}) = \sigma_\varepsilon^2 \sum_{k=1}^{n_d} (\pi_{dk}\sum_{l=1}^{n_d}\frac{1}{\pi_{dl}})^{-2} = \sigma_\varepsilon^2 \delta_{dw}.$$

**Eblup under a mixed model**

When $\sigma_v^2$ and $\sigma_\varepsilon^2$ are known, it follows from Rao (1973); Henderson (1975); Battese et al. (1988); Prasad & Rao (1990); You & Rao (2002); Rao (2003) that the best linear unbiased predictor (BLUP) of $\mu$ is given by

$$\tilde{\mu}_d = \gamma_d \overline{y}_d + (\overline{\mathbf{X}}_d - \gamma_d \overline{\mathbf{x}}_d)^\top \tilde{\boldsymbol{\beta}}, \tag{3.11}$$

where $\tilde{\boldsymbol{\beta}}$ is the generalized least square estimator of $\boldsymbol{\beta}$,

$$\overline{y}_d = \frac{1}{n_d} \sum_{k=1}^{n_d} y_{dk},$$

$$\overline{\mathbf{x}}_d = \frac{1}{n_d} \sum_{k=1}^{n_d} \mathbf{x}_{dk}$$

and

$$\gamma_d = \left(1 + \frac{\sigma_\varepsilon^2}{\sigma_v^2 n_d}\right)^{-1}.$$

The variances $\sigma_v^2$ and $\sigma_\varepsilon^2$ can be estimated using two ordinary least squares regressions and the method of moments (see, for instance, Fuller & Battese, 1973; You & Rao, 2002; Rao, 2003).

The expression of $\tilde{\mu}_d$ in (3.11) can be re-written as

$$\tilde{\mu}_d = \gamma_d \left[\overline{y}_d + (\overline{\mathbf{X}}_d - \overline{\mathbf{x}}_d)^\top \tilde{\boldsymbol{\beta}}\right] + (1 - \gamma_d) \overline{\mathbf{X}}_d^\top \tilde{\boldsymbol{\beta}}.$$

The empirical best linear unbiased predictor estimator (EBLUP) of $\tilde{\mu}_d$, denoted $\hat{\mu}_d$, is then obtained by replacing $\sigma_v^2$ and $\sigma_\varepsilon^2$ by $\hat{\sigma}_v^2$ and $\hat{\sigma}_\varepsilon^2$ in the expression of $\gamma_d$

$$\hat{\mu}_d = \hat{\gamma}_d \left[\overline{y}_d + (\overline{\mathbf{X}}_d - \overline{\mathbf{x}}_d)^\top \hat{\boldsymbol{\beta}}\right] + (1 - \hat{\gamma}_d) \overline{\mathbf{X}}_d^\top \hat{\boldsymbol{\beta}},$$

where $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2)$.

It follows that the expression of the total estimator is obtained by

$$\begin{aligned} \hat{t}_{y,\text{EBLUP}}^d &= N_d \hat{\mu}_d, \\ &= \hat{\gamma}_d \left(N_d \overline{y}_d + (\mathbf{t}_x^d - N_d \overline{\mathbf{x}}_d)^\top \hat{\boldsymbol{\beta}}\right) + (1 - \hat{\gamma}_d) \mathbf{t}_x^{d\top} \hat{\boldsymbol{\beta}}, \\ &= \hat{\gamma}_d \hat{t}_{y,\text{EBLUP,direct}}^d + (1 - \hat{\gamma}_d) \hat{t}_{y,\text{EBLUP,synth}}^d \end{aligned} \tag{3.12}$$

**Pseudo-Eblup under a mixed model**

The pseudo-BLUP estimator is obtained from a combination of the model (3.9) with the sample weights. From assuming that $\sigma_v^2$ and $\sigma_\varepsilon^2$ are known,

the pseudo-BLUP estimator of $\mu_d$ from the aggregated model (3.10) is given by

$$\tilde{\mu}_{dw} = \gamma_{dw}\left[\overline{y}_{dw} + (\overline{\mathbf{X}}_{dw} - \overline{\mathbf{x}}_{dw})^\top \tilde{\boldsymbol{\beta}}_w\right] + (1-\gamma_{dw})\overline{\mathbf{X}}_{dw}\tilde{\boldsymbol{\beta}}_w, \quad (3.13)$$

where

$$\gamma_{dw} = \left(1 + \frac{\sigma_\varepsilon^2 \delta_{dw}}{\sigma_v^2}\right)^{-1}$$

and

$$\tilde{\boldsymbol{\beta}}_w = \left[\sum_{d=1}^{D}\sum_{k=1}^{n_d} \frac{\mathbf{x}_{dk}}{\pi_{dk}}(\mathbf{x}_{dk} - \gamma_{dw}\overline{\mathbf{x}}_{dw})^\top\right]^{-1}\left[\sum_{d=1}^{D}\sum_{k=1}^{n_d}(\mathbf{x}_{dk} - \gamma_{dw}\overline{\mathbf{x}}_{dw})\frac{y_{dk}}{\pi_{dk}}\right].$$

The pseudo-empirical best linear unbiased predictor estimator (pseudo-EBLUP) of $\tilde{\mu}_{dw}$, denoted $\hat{\mu}_{dw}$, is then obtained by replacing $\sigma_v^2$ and $\sigma_\varepsilon^2$ by $\hat{\sigma}_v^2$ and $\hat{\sigma}_\varepsilon^2$

$$\hat{\mu}_{dw} = \hat{\gamma}_{dw}\left[\overline{y}_{dw} + (\overline{\mathbf{X}}_{dw} - \overline{\mathbf{x}}_{dw})^\top \hat{\boldsymbol{\beta}}_w\right] + (1-\hat{\gamma}_{dw})\overline{\mathbf{X}}_{dw}\hat{\boldsymbol{\beta}}_w,$$

where $\hat{\boldsymbol{\beta}}_w = \tilde{\boldsymbol{\beta}}_w(\hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2)$.

It also follows that the expression of the total estimator is obtained by

$$\begin{aligned}
\hat{t}^d_{y,\text{p-EBLUP}} &= N_d\hat{\mu}_{dw}, \\
&= \hat{\gamma}_{dw}\left(N_d\overline{y}_{dw} + (\mathbf{t}_x^d - N_d\overline{\mathbf{x}}_{dw})^\top\hat{\boldsymbol{\beta}}_w\right) + (1-\hat{\gamma}_{dw})\mathbf{t}_x^{d\top}\hat{\boldsymbol{\beta}}_w, \\
&= \hat{\gamma}_{dw}\hat{t}^d_{y,\text{p-EBLUP,direct}} + (1-\hat{\gamma}_{dw})\hat{t}^d_{y,\text{p-EBLUP,synth}} \quad (3.14)
\end{aligned}$$

For a given area $A_d$, when the weights are calibrated with the known size $N_d$ i.e.

$$\sum_{k=1}^{n_d}\frac{1}{\pi_{dk}} = N_d,$$

and when the unit level model (3.9) includes the intercept term, the estimator $\hat{t}^d_{y,\text{p-EBLUP}}$ satisfies the benchmarking property without any adjustment (see for instance You & Rao, 2002; Rao, 2003).

In the case that the weights are not calibrated with the known size $N_d$, a preliminary calibration on the weights can be done in order to obtain a coherent pseudo-EBLUP estimator. It follows that under a simple random sampling without replacement, a preliminary calibration on the weights leads to the equality $\gamma_d = \gamma_{dw}$. In fact, under a simple random sampling without replacement, the sampling weights are all equal: $1/\pi_{dk} = N/n$ for all $k$. Then, new weights calibrated on the known size $N_d$ are obtained: $1/\pi_{dk,w} = N_d/n_d$. It follows that $\delta_{dw} = \sum_{k=1}^{n_d}(\frac{N_d/n_d}{\sum_{l=1}^{n_d} N_d/n_d})^2 = 1/n_d$ and then $\gamma_d = \gamma_{dw}$.

#### Composite form of the Eblup and the pseudo-Eblup

The EBLUP and the pseudo-EBLUP can be seen as composite estimators. They are weighted averages of a regression synthetic estimator and a pseudo-direct estimator. Similarly to the proposed method, when $n_d$ tends to 0 (resp. $+\infty$), $\gamma_d$ and $\gamma_{dw}$ tend to 0 (resp. $+\infty$): when the size of an area is large enough, more weight is attached to the direct estimation part and vice versa. Then, an estimation of the tuning constant $\alpha_d$ can be obtained using an analogy with the parameters $\gamma_d$ and $\gamma_{dw}$ of the BLUP:

$$\hat{\alpha}_d^{\{1\}} = 1 - \hat{\gamma}_d, \tag{3.15}$$

or

$$\hat{\alpha}_d^{\{2\}} = \hat{\alpha}_{dw} = 1 - \hat{\gamma}_{dw}. \tag{3.16}$$

## 3.7 Simulation study

### 3.7.1 Simulated data with a mixed model

In order to test the performance of the proposed methodology, we ran a set of simulations.

#### Simulated population

A population of $N = 2,000$ units, with $D = 20$ disjoint areas of sizes $(N_1, \ldots, N_D)$, is created from a linear mixed model given in (3.9):

$$\begin{aligned}
\mathbf{x}_{dk} &= \begin{pmatrix} 1 & x_{dk} \end{pmatrix}^\top \text{ with } x_{dk} \overset{\text{iid}}{\sim} \mathcal{N}(20, \sigma_x^2 = 9), \\
\beta &= \begin{pmatrix} 12 & 0.4 \end{pmatrix}^\top, \\
v_d &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_v^2 = 4), \\
\varepsilon_{dk} &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2 = 1).
\end{aligned}$$

The Figure 3.1 gives an overview of a generated population.

#### Precision comparison between the weights splitting estimator and some classical small area estimators

From the generated population, $B = 10,000$ samples of size $n = 200$ are drawn by a simple random sampling without replacement. Within areas, the EBLUP estimator, the pseudo-EBLUP estimator and the proposed estimator (with their respective direct and synthetic components) are computed to estimate the totals $t_y^d$, for $d = 1, \ldots, 20$. The relative root mean square error (%RRMSE) is used to quantify the performance of the estimators.

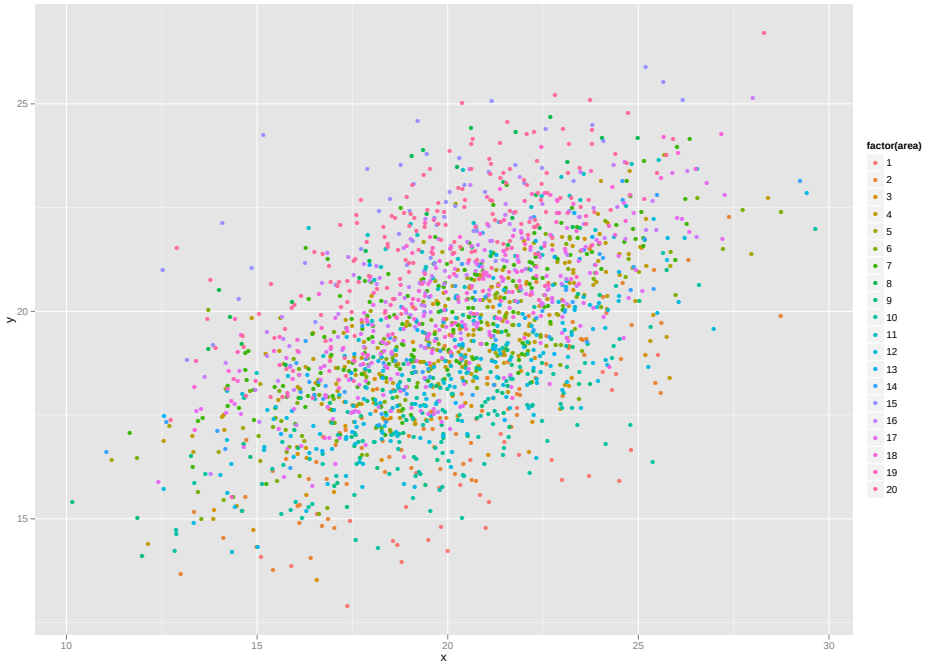


Figure 3.1 – *Generated population from the linear mixed model given in 3.7.1: the different colors indicate the areas).*

For a given estimator $\hat{t}_y^d$, the %RRMSE$_d$ is obtained as follows

$$\%\text{RRMSE}_d = 100 \times \frac{\sqrt{\text{MSE}_d}}{t_y^d}, \tag{3.17}$$

where MSE$_d$ is the sum of the square of the bias and the variance

$$\text{MSE}_d = \left( \frac{1}{B} \sum_{b=1}^{B} \hat{t}_{y,b}^d - t_y^d \right)^2 + \frac{N-n}{N-1} \frac{1}{B-1} \sum_{b=1}^{B} \left( \hat{t}_{y,b}^d - \frac{1}{B} \sum_{b=1}^{B} \hat{t}_{y,b}^d \right)^2 .$$

Table 3.1 shows a comparison between the proposed method and the Eblup estimator. Table 3.2 shows a comparison between the proposed method and the pseudo-Eblup estimator. For each simulation run, each size $n_d$ (for $d = 1, \ldots, D$) is not fixed because the sampling design is a simple random sampling without replacement from the overall population. That is why, the second columns of Table 3.1 and Table 3.2 represent the means $\overline{n}_d$ (for $d = 1, \ldots, D$) of each area size through the $B = 10,000$ drawings. The tuning constants $\hat{\alpha}_d$ and $\hat{\alpha}_{dw}$ also are re-estimated in each simulation run. That also is why, the third columns of Table 3.1 and Table 3.2 represent the means $\overline{\hat{\alpha}}_d$ (resp. $\overline{\hat{\alpha}}_{wd}$) (for $d = 1, \ldots, D$) of each area tuning constant through the $B = 10,000$ drawings. As discussed in Section 3.6.2, the equalities $\hat{\alpha}_d = \hat{\alpha}_{wd}$ and $\overline{\hat{\alpha}}_d = \overline{\hat{\alpha}}_{wd}$ are obtained because the sampling design is a simple random sampling without replacement from the overall population.

Table 3.1 – *Computed %RRMSE of the* Eblup *estimator, of the proposed estimator, and their respective component estimators, from 10,000 drawn samples of size 200 by a simple random sampling without replacement from the generated population (see Figure 3.1).*

| | | | Eblup components | | | Weights Splitting method | | |
|---|---|---|---|---|---|---|---|---|
| Area | $\overline{n}_d$ | $\overline{\hat{\alpha}}_d$ | Direct | Synthetic | Eblup | Calibration | Extra | Composite |
| 1 | 3.45 | 0.14 | 5.26 | 24.68 | 5.95 | 49.14 | 24.10 | 9.70 |
| 2 | 10.03 | 0.04 | 1.94 | 10.62 | 1.92 | 27.72 | 10.09 | 2.96 |
| 3 | 5.66 | 0.07 | 2.68 | 8.65 | 2.55 | 37.40 | 8.14 | 3.24 |
| 4 | 17.40 | 0.02 | 1.10 | 1.34 | 1.07 | 20.50 | 0.67 | 0.98 |
| 5 | 4.69 | 0.09 | 2.58 | 1.48 | 2.32 | 41.11 | 1.52 | 1.91 |
| 6 | 12.53 | 0.03 | 1.29 | 3.60 | 1.26 | 25.07 | 3.03 | 1.31 |
| 7 | 18.67 | 0.02 | 1.09 | 2.03 | 1.07 | 19.91 | 2.20 | 1.06 |
| 8 | 2.81 | 0.17 | 3.94 | 11.88 | 4.02 | 53.72 | 12.22 | 5.53 |
| 9 | 7.28 | 0.05 | 2.02 | 7.82 | 1.96 | 32.72 | 7.31 | 2.62 |
| 10 | 15.01 | 0.02 | 1.41 | 12.35 | 1.41 | 22.41 | 11.83 | 2.61 |
| 11 | 4.41 | 0.10 | 2.20 | 7.38 | 2.14 | 43.12 | 7.73 | 3.05 |
| 12 | 18.15 | 0.02 | 1.14 | 7.11 | 1.12 | 20.45 | 6.59 | 1.59 |
| 13 | 3.51 | 0.14 | 2.97 | 1.10 | 2.57 | 48.92 | 0.71 | 2.08 |
| 14 | 7.76 | 0.05 | 1.76 | 1.11 | 1.67 | 31.96 | 0.65 | 1.46 |
| 15 | 4.88 | 0.09 | 2.46 | 14.11 | 2.63 | 40.50 | 14.44 | 5.08 |
| 16 | 12.40 | 0.03 | 1.28 | 5.36 | 1.25 | 24.86 | 5.69 | 1.60 |
| 17 | 16.82 | 0.02 | 1.22 | 1.62 | 1.19 | 21.13 | 0.95 | 1.12 |
| 18 | 11.60 | 0.03 | 1.25 | 3.65 | 1.22 | 25.65 | 3.95 | 1.39 |
| 19 | 6.48 | 0.06 | 1.84 | 5.57 | 1.76 | 35.13 | 5.89 | 2.19 |
| 20 | 16.46 | 0.02 | 1.08 | 10.76 | 1.10 | 21.53 | 11.11 | 2.23 |

In Table 3.1, the direct estimation component of the Eblup estimator appears to be generally better than the global calibration which is the direct estimation part of the weights splitting method. This can be explained by the fact that the direct estimation component of the Eblup is a regression estimator built from the parameter $\hat{\beta}$ of the synthetic estimation component of the Eblup. The synthetic part of the Eblup and the extra-contribution part seem to have equivalent performance. In spite of the weakness of the direct estimation part of the proposed composite estimator relative to the direct estimation part of the Eblup, the two composite estimators perform equivalently thanks to the matrix calibration computed during the weights splitting method procedure. The result shown in Table 3.1 is quite difficult to interpret, because the weight-splitting estimator incorporates sampling weights and is benchmarked, while the Eblup estimator does not.

Table 3.2 seems to be fairer in terms of comparison because both the pseudo-Eblup estimator and the proposed estimator respect the benchmarking property and use sampling weights. As previously, whereas the synthetic part of the pseudo-Eblup and of the proposed method seem to perform equivalently, the direct estimation component of the pseudo-

Table 3.2 – *Computed %RRMSE of the pseudo-*Eblup *estimator, of the proposed estimator, and their respective component estimators, from 10,000 drawn samples of size 200 by a simple random sampling without replacement from the generated population (see Figure 3.1).*

| | | | pseudo-Eblup components | | | Weights Splitting method | | |
|---|---|---|---|---|---|---|---|---|
| Area | $\overline{n}_d$ | $\overline{\hat{\alpha}}_{dw}$ | Direct | Synthetic | pseudo-Eblup | Calibration | Extra | Composite |
| 1 | 3.45 | 0.14 | 5.25 | 24.58 | 5.97 | 49.14 | 24.10 | 9.70 |
| 2 | 10.03 | 0.04 | 1.94 | 10.49 | 1.92 | 27.72 | 10.09 | 2.96 |
| 3 | 5.66 | 0.07 | 2.67 | 8.51 | 2.54 | 37.40 | 8.14 | 3.24 |
| 4 | 17.40 | 0.02 | 1.10 | 0.86 | 1.07 | 20.50 | 0.67 | 0.98 |
| 5 | 4.69 | 0.09 | 2.58 | 1.16 | 2.30 | 41.11 | 1.52 | 1.91 |
| 6 | 12.53 | 0.03 | 1.29 | 3.38 | 1.26 | 25.07 | 3.03 | 1.31 |
| 7 | 18.67 | 0.02 | 1.09 | 1.83 | 1.07 | 19.91 | 2.20 | 1.06 |
| 8 | 2.81 | 0.17 | 3.94 | 11.90 | 4.03 | 53.72 | 12.22 | 5.53 |
| 9 | 7.28 | 0.05 | 2.02 | 7.68 | 1.96 | 32.72 | 7.31 | 2.62 |
| 10 | 15.01 | 0.02 | 1.41 | 12.23 | 1.41 | 22.41 | 11.83 | 2.61 |
| 11 | 4.41 | 0.10 | 2.20 | 7.38 | 2.14 | 43.12 | 7.73 | 3.05 |
| 12 | 18.15 | 0.02 | 1.14 | 6.97 | 1.12 | 20.45 | 6.59 | 1.59 |
| 13 | 3.51 | 0.14 | 2.97 | 0.49 | 2.55 | 48.92 | 0.71 | 2.08 |
| 14 | 7.76 | 0.05 | 1.76 | 0.49 | 1.66 | 31.96 | 0.65 | 1.46 |
| 15 | 4.88 | 0.09 | 2.47 | 14.14 | 2.64 | 40.50 | 14.44 | 5.08 |
| 16 | 12.40 | 0.03 | 1.28 | 5.33 | 1.25 | 24.86 | 5.69 | 1.60 |
| 17 | 16.82 | 0.02 | 1.22 | 1.22 | 1.19 | 21.13 | 0.95 | 1.12 |
| 18 | 11.60 | 0.03 | 1.25 | 3.58 | 1.22 | 25.65 | 3.95 | 1.39 |
| 19 | 6.48 | 0.06 | 1.84 | 5.55 | 1.76 | 35.13 | 5.89 | 2.19 |
| 20 | 16.46 | 0.02 | 1.08 | 10.78 | 1.10 | 21.53 | 11.11 | 2.23 |

Eblup performs better than the direct estimation part of the weights splitting method. This also can be explained by the construction of the direct estimation of the pseudo-Eblup with the parameter $\hat{\beta}_w$ which is derived from the synthetic part of the pseudo-Eblup. Despite this disadvantage, when the calibration and extra-contribution parts of the weights splitting method are mixed and are re-calibrated on the rows (benchmarking constraints) and on the columns (calibration constraints of the areas), the pseudo-Eblup estimator and the weights splitting estimator appear to perform equivalently. These obtained similar efficiencies can also partly be explained by the same weight attached to the synthetic component of the pseudo-Eblup and attached to the extra-contribution component of the proposed composite estimator.

**Resampling procedure to estimate the variance of the weights splitting estimator**

From the generated artificial population (see Figure 3.1), a resampling procedure can be performed to estimate the variance of the weights splitting estimator. Given a sample drawn from the generated population, $B$ samples are drawn from this initial sample (for instance $B = 500$). The $B$

weights splitting estimators are computed from these $B$ new samples. The variance by bootstrap is obtained from computing the empirical variance of the $B$ weights splitting estimators. The considered sampling design is always a simple random sampling without replacement.

In order to test the efficiency of the variance estimation by bootstrap, the Algorithm 3.1 is considered. In the Algorithm 3.1, each sample $s^{b^*}$

**Algorithm 3.1:** Resampling procedure for testing the efficiency of the variance estimation by bootstrap

```
Consider the generated artificial population (see Figure 3.1);
Consider B* = B** = 500 the number of iterations;
Consider D = 20 the number of areas;
Consider n = 200 the sample size;
for each iteration b* do
    Draw a sample s^{b*} of size n by a simple random sampling
    without replacement from the generated population;
    Compute the weights splitting estimators
    (t̂^1_{y,c})^{b*}, ..., (t̂^d_{y,c})^{b*}, ..., (t̂^D_{y,c})^{b*};
    for each iteration b** do
        Draw a sample s^{b*}_{b**} of size n = 200 by a simple random
        sampling with replacement from s^{b*};
        Compute the weights splitting estimators
        (t̂^1_{y,c})^{b*}_{b**}, ..., (t̂^d_{y,c})^{b*}_{b**}, ..., (t̂^D_{y,c})^{b*}_{b**}
    Compute (σ̂^{b*}_1, ..., σ̂^{b*}_d, ..., σ̂^{b*}_D), the empirical variances of the
    B** obtained weights splitting estimators
    (t̂^1_{y,c})^{b*}_{b**}, ..., (t̂^d_{y,c})^{b*}_{b**}, ..., (t̂^D_{y,c})^{b*}_{b**}
Compute the empirical variances of the B* weights splitting
estimators (t̂^1_{y,c})^{b*}, ..., (t̂^d_{y,c})^{b*}, ..., (t̂^D_{y,c})^{b*};
Compute the empirical means of the B* variances
(σ̂^{b*}_1, ..., σ̂^{b*}_d, ..., σ̂^{b*}_D) (bootstrap variance);
```

drawn from the generated population leads to weights splitting estimators $(\hat{t}^d_{y,c})^{b^*}$ with bootstrap variances $\hat{\sigma}^{b^*}_d$ for $d = 1 \dots D$, where

$$\begin{aligned}\hat{\sigma}^{b^*}_d &= \mathrm{Var}[(\hat{t}^d_{y,c})^{b^*}_{b^{**}}] \\ &= \frac{N-n}{N-1}\frac{1}{B^{**}-1}\sum_{b^{**}=1}^{B^{**}}\left((\hat{t}^d_{y,c})^{b^*}_{b^{**}} - \frac{1}{B^{**}}\sum_{b^{**}=1}^{B^{**}}(\hat{t}^d_{y,c})^{b^*}_{b^{**}}\right)^2.\end{aligned}$$

For $d = 1 \dots D$, for $b^* = 1 \dots B^*$, consider

$$\mathrm{Var}[(\hat{t}^d_{y,c})^{b^*}] = \frac{N-n}{N-1}\frac{1}{B^*-1}\sum_{b^*=1}^{B^*}\left((\hat{t}^d_{y,c})^{b^*} - \frac{1}{B^*}\sum_{b^*=1}^{B^*}(\hat{t}^d_{y,c})^{b^*}\right)^2$$

Table 3.3 – *Variance of the weights splitting estimators* vs *Mean of the bootstrap variances of the weights splitting estimators.*

| Area | $\mathrm{Var}[(\hat{t}^d_{y,c})^{b^*}]$ | $\tilde{\mathbb{E}}_{sim}\{\hat{\sigma}^{b^*}_d\}$ |
|---|---|---|
| 1 | 1264.0 | 989.3 |
| 2 | 1566.2 | 1550.0 |
| 3 | 612.8 | 570.0 |
| 4 | 1021.0 | 1163.5 |
| 5 | 324.6 | 274.9 |
| 6 | 778.4 | 810.6 |
| 7 | 1288.8 | 1573.6 |
| 8 | 574.2 | 432.9 |
| 9 | 713.8 | 699.0 |
| 10 | 2555.2 | 2768.5 |
| 11 | 505.2 | 409.1 |
| 12 | 2088.3 | 1981.9 |
| 13 | 218.0 | 164.5 |
| 14 | 485.1 | 472.0 |
| 15 | 1442.2 | 1264.0 |
| 16 | 1134.5 | 1290.5 |
| 17 | 1275.3 | 1383.5 |
| 18 | 813.8 | 837.5 |
| 19 | 528.9 | 543.1 |
| 20 | 3438.0 | 3849.4 |

the simulated variance of the weights splitting estimator and consider

$$\tilde{\mathbb{E}}_{sim}\{\hat{\sigma}^{b^*}_d\} \quad = \quad \frac{1}{B^*}\sum_{b^*=1}^{B^*}\hat{\sigma}^{b^*}_d$$

the simulated expectation of the weights splitting bootstrap variance estimator. Table 3.3 gives a comparison between these two quantities. It shows that the two quantities are very closed.
Once the bootstrap variances computed, construction of confidence intervals can be processed. Table 3.4 reports the numbers of times (%) the true total values $t^d_y$ (for $d = 1 \ldots D$), from the generated population, lie within the 95% confidence intervals built with bootstrap variances. Table 3.4 shows that given an area, the number of times the true total value lies within the 95 % confidence interval increases with the size of the area.

### 3.7.2 Data with county crop areas

We compare the pseudo-Eblup and the weights splitting procedures applied to a real data given by Battese et al. (1988) and taken up by You

Table 3.4 – *Numbers of times (%) the true total values $t_y^d$ (for $d = 1 \dots D$), from the generated population (see Figure 3.1), lie within the* 95% *confidence intervals built with bootstrap variances.*

| Area | $N_d$ | Numbers of times (%) |
|---|---|---|
| 1 | 34 | 75 |
| 2 | 100 | 84 |
| 3 | 57 | 78 |
| 4 | 174 | 93 |
| 5 | 47 | 84 |
| 6 | 125 | 91 |
| 7 | 187 | 93 |
| 8 | 28 | 65 |
| 9 | 73 | 85 |
| 10 | 150 | 87 |
| 11 | 44 | 75 |
| 12 | 182 | 87 |
| 13 | 35 | 88 |
| 14 | 78 | 91 |
| 15 | 49 | 73 |
| 16 | 124 | 89 |
| 17 | 168 | 94 |
| 18 | 116 | 87 |
| 19 | 65 | 86 |
| 20 | 164 | 87 |

& Rao (2002); Rao (2003). These authors wanted to estimate the mean of hectares of corn per segment for $D = 12$ counties in north-central Iowa. Each county is divided into area segments and the areas under corn are in the area segments. The authors used a sample $s$ of $n = 36$ segments and assumed simple random sampling within areas. The population is assumed to follow the linear mixed model given in (3.9) where the function of interest is the number of hectares of corn per segment per county, and the auxiliary information is the number of pixels seen as corn and as soybeans.

Table 3.5 reports the pseudo-Eblup and weights splitting estimates of hectares of corn with their respective coefficients of variation. The variances are obtained by a resampling procedure of 1,000 iterations (see Algorithm 3.2). The two estimators automatically respect the benchmarking property. Indeed, we have $\sum_{d=1}^{D} N_d \hat{\mu}_{dw} = \hat{\mathbf{t}}_{y,\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x,\pi})^\top \hat{\boldsymbol{\beta}}_w = 815016,3$ and $\sum_{d=1}^{D} \hat{t}_{y,c}^d = \hat{t}_{y,w} = 817087,2$. Table 3.5 gives quite similar estimates with similar efficiencies.

Table 3.5 – *Estimated hectares of corn with coefficients of variation.*

| County | $n_d$ | pseudo-Eblup | | Weights Splitting | |
|---|---|---|---|---|---|
| | | Estimate | c.v. (%) | Estimate | c.v. (%) |
| | | Corn | | | |
| Cerro Gordo | 1 | 120.5 | 2.6 | 121.8 | 3.0 |
| Hamilton | 1 | 125.3 | 2.7 | 122.7 | 3.2 |
| Worth | 1 | 106.3 | 7.8 | 108.3 | 6.4 |
| Humboldt | 2 | 107.3 | 8.4 | 111.1 | 5.7 |
| Franklin | 3 | 143.8 | 4.5 | 142.8 | 5.3 |
| Pocahontas | 3 | 111.5 | 5.0 | 111.8 | 6.0 |
| Winnebago | 3 | 112.1 | 5.5 | 113.8 | 4.6 |
| Wright | 3 | 121.3 | 3.8 | 120.2 | 3.3 |
| Webster | 4 | 115.1 | 3.3 | 114.7 | 4.5 |
| Hancock | 5 | 124.5 | 3.4 | 124.2 | 3.1 |
| Kossuth | 5 | 106.6 | 3.2 | 109.3 | 3.8 |
| Hardin | 5 | 143.5 | 3.3 | 141.0 | 3.0 |

Source: LandSat data from Table 1 in Battese et al. (1988, p. 29).

**Algorithm 3.2:** Resampling procedure for the variance estimation by bootstrap

1 Consider $B = 1,000$ the number of iterations;
2 Consider $n = 36$ the sample size;
3 **for** *each iteration* $b$ **do**
4 Draw a sample $s^b$ of size $n$ by a simple random sampling with replacement within areas from the sample $s$ (the area sample sizes are fixed and are the same as those in $s$);
5 Compute the pseudo-Eblup and weights splitting estimators from the sample $s^b$
6 Compute the variances of the $B$ obtained pseudo-Eblup and weights splitting estimators with considering the finite population correction factor;

## 3.8 Concluding remarks

The simulations show that, even when the data are really generated by a mixed model, the proposed estimator does not seem worse than the Eblup and pseudo-Eblup estimators. The proposed estimators offer several advantages. The regional estimates are coherent with the overall estimates. The estimator takes the sampling weights into account. The estimator can be written as a weighting system and can thus be applied on any variable of interest.

As for all the composite estimators, the proposed estimator clearly states that the parameters $\alpha_d$ (for $d = 1 \dots D$) depend on the variable of inter-

est. In the case of a set of variables of interest belonging to a specified theme, the constant $\alpha_d$ can be chosen as the mean of the tuning constants obtained from each variable of interest.

The next step will consist of computing a variance estimator of the composite estimator. Since the proposed estimator is obtained by successive matrix calibrations, the computation of this variance is complex, which is thus a challenging objective. If a closed form of variance estimator is intractable, resampling procedures for simple random sampling as seen in Section 3.7.1 can be considered; resampling procedures for sampling design with unequal inclusion probabilities can be based on Antal & Tillé (2011) methodology.

The proposed method is not particularly robust for the resistance to outliers. It could be an interesting further topic of research.

## Acknowledgments

The authors would like to thank an anonymous referee for useful comments and suggestions.

# 4 Small area estimators from a synthetic population generated by a nested error regression model

**Abstract** A fixed synthetic population is assumed to be generated by a nested error regression model with equal errors. A sample is drawn from the population following a given sampling design and small area mean estimators are obtained. These estimators are calibrated on the auxiliary variables and are homogeneous linear. The obtained estimators can be rewritten as weighted sums. The weights depend on the considered variable of interest. Given a set of variables of interest belonging to a specified theme, a discussion is lead on building a unique weighting system and on the possibility of applying it on new variables.

**Keywords** indirect estimation, nested error regression model, design-based inference, weights

## 4.1 Introduction

Under a nested error regression model, small areas are contained in the random part of the linear mixed model. In order to estimate small area parameters, two different forms of the Best Linear Unbiased Predictor (BLUP) have been respectively proposed by Royall (1970) and by Henderson (1975). These two estimators are definitely not equal (see, for instance, Guggemos & Tillé, 2009). In this paper, we are only going to consider the form of the BLUP proposed by Henderson (1975). The Empirical Best Linear Unbiased Predictor (EBLUP) is an estimator derived from the BLUP from replacing some parameters with theirs respective estimators. Another derived estimator, called pseudo-EBLUP, has been proposed by You & Rao (2002), which combines the nested error regression model with the

use of the sampling weights. This new estimator has the advantage to automatically respect the benchmarking property.

A synthetic population is generated by a nested error regression model with equal errors, i.e. with a homoscedastic property. When generated, the population is then considered as fixed. One of the goal of this paper is, from the synthetic population, to rewrite the Eblup and the pseudo-Eblup estimators into a homogenous linear form. Weights can then be extracted from these new forms but obviously depend on the considered variable of interest. Given a set of variables of interest, enough correlated with each other, a discussion is lead on choosing a unique weighting system. Next, given another variable of interest, also correlated with the previous set of variables, a discussion is lead on a transferred weighting system.

The paper is structured as follows. Section 4.2 gives a general framework of the model used to generate the population. Sections 4.3 and 4.4 recall the theorical notion of the construction of the Eblup and of the pseudo-Eblup estimators. A unique compromise weigthing system and a transferred weighting system are respectively introduced in Section 4.5 and in Section 4.6. Section 4.7 presents a set of simulations, and the paper ends with some concluding remarks in Section 4.8.

## 4.2 Population model

Consider a finite population $U$ of size $N$ divided into $m$ disjoint areas $U_i$ for $i = 1, \dots, m$, of sizes $N_i$ for $i = 1, \dots, m$. We are interested in estimating area characteristics. Consider the variable of interest $y_{ij}$ for the $j$-th unit within the $i$-th area which is obtained from a nested error regression model (Henderson, 1975; Battese et al., 1988; Prasad & Rao, 1990; You & Rao, 2002; Rao, 2003) as follows

$$y_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + v_i + \varepsilon_{ij} \tag{4.1}$$

where $j = 1, \dots, N_i$, $i = 1, \dots, m$, $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp})^\top$ are vectors of available auxiliary variables, $v_i$ are independent centered normal variables with variances $\sigma_v^2$, $\varepsilon_{ij}$ are independent centered normal variables with variances $\sigma_\varepsilon^2$. Moreover, the $v_i$ are assumed to be independent from the $\varepsilon_{ij}$. The parameters of interest are the area means $\overline{y}_i = N^{-1} \sum_{j=1}^{N_i}$ for $i = 1, \dots, m$.

A sample $s$ of size $n$ is drawn from the fixed population $U$. The subsample $s_i$ is the subsample from area $U_i$ i.e. $s_i = s \cap U_i$, of size $n_i$, for $i = 1, \dots, m$, where $n = \sum_{i=1}^m n_i$. The sample data are assumed to be

obtained from the following model

$$y_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + v_i + \varepsilon_{ij} \tag{4.2}$$

where $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$.

## 4.3 EBLUP estimator

### 4.3.1 Theorical reminder

The model (4.2) can be re-written in a matrix notation as follows

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + v_i + \boldsymbol{\varepsilon}_i \tag{4.3}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$, $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})^\top$, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^\top$, and $\text{Var}(\mathbf{y}_i) = \mathbf{V}_i = \sigma_v^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^\top + \sigma_\varepsilon^2 \mathbf{I}_{n_i}$, for $i = 1, \ldots, m$, $\mathbf{1}_k$ denoting a vector of ones of size $k$ and $\mathbf{I}_k$ denoting the identity matrix of size $k \times k$.

The log-likelihood is then given by

$$\mathcal{L}(\boldsymbol{\beta}, \sigma_v^2, \sigma_\varepsilon^2, \mathbf{y}_i) = c - \frac{1}{2} \sum_{i=1}^{m} \{\log |\mathbf{V}_i| + (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})^\top \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})\}. \tag{4.4}$$

We obtain

$$\mathbf{V}_i^{-1} = \sigma_\varepsilon^{-2} \left( \mathbf{I}_{n_i} - \frac{\gamma_i}{n_i} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^\top \right)$$

where

$$\gamma_i = \left( 1 + \frac{\sigma_\varepsilon^2}{\sigma_v^2 n_i} \right)^{-1}, \tag{4.5}$$

from using the following matrix inversion formula (see, for instance, Rao, 2003, p. 135)

$$(A + \mathbf{u}\mathbf{v}^\top)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^\top A^{-1}}{1 + \mathbf{v}^\top A^{-1} \mathbf{u}}.$$

The log-likelihood (4.4) can be then written as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}, \sigma_v^2, \sigma_\varepsilon^2, \mathbf{y}_i) \quad = \quad & c - \frac{1}{2} \sum_{i=1}^{m} \Bigg[ \log |\mathbf{V}_i| + \sigma_\varepsilon^{-2} \sum_{j=1}^{n_i} [(y_{ij} - \mathbf{x}_{ij}^\top \boldsymbol{\beta})^2 \\ & - \gamma_i (y_{ij} - \mathbf{x}_{ij}^\top \boldsymbol{\beta})^2] \Bigg]. \end{aligned} \tag{4.6}$$

By setting the derivative of (4.6) with respect to $\boldsymbol{\beta}$ to zero, given by

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbf{x}_{ij} [y_{ij} - \mathbf{x}_{ij}^\top \boldsymbol{\beta} - \gamma_i (\overline{y}_i - \overline{\mathbf{x}}_i^\top \boldsymbol{\beta})] = 0 \tag{4.7}$$

where $\overline{y}_i = {n_i}^{-1} \sum_{j=1}^{n_i} y_{ij}$ and $\overline{\mathbf{x}}_i = n_i^{-1} \sum_{j=1}^{n_i} \mathbf{x}_{ij}$, we obtain the value for $\boldsymbol{\beta}$ that maximizes (4.6):

$$\tilde{\boldsymbol{\beta}} = \left[ \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbf{x}_{ij} (\mathbf{x}_{ij} - \gamma_i \overline{\mathbf{x}}_i)^\top \right]^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \gamma_i \overline{\mathbf{x}}_i) y_{ij}. \tag{4.8}$$

When $\sigma_v^2$ and $\sigma_\varepsilon^2$ are known, it follows from Rao (1973); Henderson (1975); Battese et al. (1988); Prasad & Rao (1990); You & Rao (2002); Rao (2003) that the best linear unbiased predictor (BLUP) of $\mu_i = \overline{\mathbf{X}}_i^\top \boldsymbol{\beta} + v_i$ is given by

$$\tilde{\mu}_i = \gamma_i \left[ \overline{y}_i + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_i)^\top \tilde{\boldsymbol{\beta}} \right] + (1 - \gamma_i) \overline{\mathbf{X}}_i^\top \tilde{\boldsymbol{\beta}}, \tag{4.9}$$

where $\overline{\mathbf{X}}_i = {N_d}^{-1} \sum_{j=1}^{N_i} \mathbf{x}_{ij}$.

The variances $\sigma_v^2$ and $\sigma_\varepsilon^2$ can be estimated using two ordinary least squares regressions and the method of moments (see, for instance, Fuller & Battese, 1973; You & Rao, 2002; Rao, 2003). The empirical best linear unbiased predictor estimator (EBLUP) of $\tilde{\mu}_i$, denoted $\hat{\mu}_i$, is then obtained by replacing $\sigma_v^2$ and $\sigma_\varepsilon^2$ by $\hat{\sigma}_v^2$ and $\hat{\sigma}_\varepsilon^2$ in the expression of $\gamma_i$

$$\hat{\mu}_i = \hat{\gamma}_i \left[ \overline{y}_i + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_i)^\top \hat{\boldsymbol{\beta}} \right] + (1 - \hat{\gamma}_i) \overline{\mathbf{X}}_i^\top \hat{\boldsymbol{\beta}}, \tag{4.10}$$

where $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2)$.

### 4.3.2 Reformulation of the EBLUP estimator

The EBLUP estimator is clearly composite and can be written as a weighted sum of two estimators: the survey regression estimator $\hat{\mu}_{i,dir}$ and the regression synthetic estimator $\hat{\mu}_{i,syn}$.

$$\begin{aligned} \hat{\mu}_i &= \hat{\gamma}_i \left[ \overline{y}_i + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_i)^\top \hat{\boldsymbol{\beta}} \right] + (1 - \hat{\gamma}_i) \overline{\mathbf{X}}_i^\top \hat{\boldsymbol{\beta}}, \\ &= \hat{\gamma}_i \hat{\mu}_{i,dir} + (1 - \hat{\gamma}_i) \hat{\mu}_{i,syn}. \end{aligned} \tag{4.11}$$

The two component estimators are linear. In fact, from (4.8) we can write

$$\begin{aligned} \hat{\mu}_{i,dir} &= \overline{y}_i + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_i)^\top \hat{\boldsymbol{\beta}} \\ &= \sum_{j=1}^{n_i} \frac{y_{ij}}{n_i} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_i)^\top \left[ \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbf{x}_{ij} (\mathbf{x}_{ij} - \hat{\gamma}_i \overline{\mathbf{x}}_i)^\top \right]^{-1} \sum_{k=1}^{m} \sum_{j=1}^{n_k} (\mathbf{x}_{kj} - \hat{\gamma}_k \overline{\mathbf{x}}_k) y_{kj} \\ &= \sum_{j=1}^{n_i} \frac{y_{ij}}{n_i} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_i)^\top \mathbf{T}^{-1} \sum_{k=1}^{m} \sum_{j=1}^{n_k} (\mathbf{x}_{kj} - \hat{\gamma}_k \overline{\mathbf{x}}_k) y_{kj} \\ &= \sum_{k=1}^{m} \sum_{j=1}^{n_k} \left[ \frac{y_{kj}}{n_k} \mathbb{1}_{\{k=i\}} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_i)^\top \mathbf{T}^{-1} (\mathbf{x}_{kj} - \hat{\gamma}_k \overline{\mathbf{x}}_k) y_{kj} \right] \\ &= \sum_{k=1}^{m} \sum_{j=1}^{n_k} \left[ \frac{1}{n_k} \mathbb{1}_{\{k=i\}} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_i)^\top \mathbf{T}^{-1} (\mathbf{x}_{kj} - \hat{\gamma}_k \overline{\mathbf{x}}_k) \right] y_{kj} \\ &= \sum_{k=1}^{m} \sum_{j=1}^{n_k} w_{i,dir}^{kj} y_{kj}, \end{aligned} \tag{4.12}$$

and

$$\begin{aligned}\hat{\mu}_{i,syn} &= \overline{\mathbf{X}}_i^\top \left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}(\mathbf{x}_{ij}-\hat{\gamma}_i\overline{\mathbf{x}}_i)^\top\right]^{-1}\sum_{k=1}^{m}\sum_{j=1}^{n_k}(\mathbf{x}_{kj}-\hat{\gamma}_k\overline{\mathbf{x}}_k)y_{kj}\\ &= \sum_{k=1}^{m}\sum_{j=1}^{n_k}\left[\overline{\mathbf{X}}_i^\top\mathbf{T}^{-1}(\mathbf{x}_{kj}-\hat{\gamma}_k\overline{\mathbf{x}}_k)\right]y_{kj}\\ &= \sum_{k=1}^{m}\sum_{j=1}^{n_k}w_{i,syn}^{kj}y_{kj}. \qquad (4.13)\end{aligned}$$

From mixing (4.12) and (4.13), a version with weights of the Eblup estimator is obtained as follows

$$\begin{aligned}\hat{\mu}_i &= \hat{\gamma}_i\sum_{k=1}^{m}\sum_{j=1}^{n_k}w_{i,dir}^{kj}y_{kj}+(1-\hat{\gamma}_i)\sum_{k=1}^{m}\sum_{j=1}^{n_k}w_{i,syn}^{kj}y_{kj}\\ &= \sum_{k=1}^{m}\sum_{j=1}^{n_k}\left[\hat{\gamma}_i w_{i,dir}^{kj}+(1-\hat{\gamma}_i)w_{i,syn}^{kj}\right]y_{kj}\\ &= \sum_{k=1}^{m}\sum_{j=1}^{n_k}w_i^{kj}y_{kj}. \qquad (4.14)\end{aligned}$$

### 4.3.3 Calibration of the Eblup estimator

The Eblup estimator and its component estimators are calibrated on the known auxiliary variables. In fact, let us apply the Eblup weights to the auxiliary variables $\mathbf{x}_{ij} = (x_{ij1},\ldots,x_{ijp})^\top$ for $j = 1,\ldots,n_i$ and $i = 1,\ldots,m$. We then have

$$\begin{aligned}\sum_{k=1}^{m}\sum_{j=1}^{n_k}w_i^{kj}\mathbf{x}_{kj}^\top &= \sum_{k=1}^{m}\sum_{j=1}^{n_k}\left[\hat{\gamma}_i w_{i,dir}^{kj}+(1-\hat{\gamma}_i)w_{i,syn}^{kj}\right]\mathbf{x}_{kj}^\top\\ &= \hat{\gamma}_i\left[\overline{\mathbf{x}}_i+(\overline{\mathbf{X}}_i-\overline{\mathbf{x}}_i)^\top\mathbf{T}^{-1}\mathbf{T}\right]+(1-\hat{\gamma}_i)\overline{\mathbf{X}}_i^\top\mathbf{T}^{-1}\mathbf{T}\\ &= \hat{\gamma}_i\overline{\mathbf{X}}_i+(1-\hat{\gamma}_i)\overline{\mathbf{X}}_i\\ &= \overline{\mathbf{X}}_i, \qquad (4.15)\end{aligned}$$

provided that $\mathbf{T}$ is invertible.
Since calibrated, the Eblup estimator is then a precise estimator when the variable of interest is very correlated with the auxiliary variables.

## 4.4 Pseudo-Eblup

### 4.4.1 Theorical reminder

You & Rao (2002) proposed a combination of the basic unit level model (4.2) with the sampling weights. Considering $w_{ij}$ the sampling weights for $j = 1,\ldots,n_i$ and $i = 1,\ldots,m$, the following weighted area level model is obtained

$$\overline{y}_{iw} = \sum_{j=1}^{n_i} \frac{w_{ij}}{\sum_{l=1}^{n_i} w_{il}} y_{ij} = \sum_{j=1}^{n_i} \frac{w_{ij}}{w_{i.}} y_{ij} = \overline{\mathbf{x}}_{iw}^{\top} \boldsymbol{\beta} + v_i + \overline{\varepsilon}_{iw} \tag{4.16}$$

where $\mathbb{E}(\overline{\varepsilon}_{iw}) = 0$ and $\text{Var}(\overline{\varepsilon}_{iw}) = \sigma_\varepsilon^2 \sum_{j=1}^{n_i} (w_{ij}/w_{i.})^2 = \sigma_\varepsilon^2 \delta_{iw}$.

The same as before, from assuming that $\sigma_v^2$ and $\sigma_\varepsilon^2$ are known, the BLUP estimator of $\mu_i = \overline{\mathbf{X}}_i^{\top} \boldsymbol{\beta} + v_i$ from the aggregated model (4.16) is given by

$$\tilde{\mu}_{iw} \quad = \quad \gamma_{iw} \left[ \overline{y}_{iw} + (\overline{\mathbf{X}}_{iw} - \overline{\mathbf{x}}_{iw})^{\top} \tilde{\boldsymbol{\beta}}_w \right] + (1 - \gamma_{iw}) \overline{\mathbf{X}}_{iw} \tilde{\boldsymbol{\beta}}_w, \tag{4.17}$$

where

$$\gamma_{iw} = \left( 1 + \frac{\sigma_\varepsilon^2 \delta_{iw}}{\sigma_v^2} \right)^{-1} \tag{4.18}$$

and

$$\tilde{\boldsymbol{\beta}}_w = \left[ \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} \mathbf{x}_{ij} (\mathbf{x}_{ij} - \gamma_{iw} \overline{\mathbf{x}}_{iw})^{\top} \right]^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} (\mathbf{x}_{ij} - \gamma_{iw} \overline{\mathbf{x}}_{iw}) y_{ij}. \tag{4.19}$$

The parameter $\tilde{\boldsymbol{\beta}}_w$ maximizes the log-likelihood of the aggregated model (4.16). It is obtained from solving the following equation which is quite similar to the equation (4.7)

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} \mathbf{x}_{ij} [y_{ij} - \mathbf{x}_{ij}^{\top} \boldsymbol{\beta} - \gamma_{iw} (\overline{y}_{iw} - \overline{\mathbf{x}}_{iw}^{\top} \boldsymbol{\beta})] = 0. \tag{4.20}$$

The pseudo-empirical best linear unbiased predictor estimator (pseudo-EBLUP) of $\tilde{\mu}_{iw}$, denoted $\hat{\mu}_{iw}$, is then obtained by replacing $\sigma_v^2$ and $\sigma_\varepsilon^2$ by $\hat{\sigma}_v^2$ and $\hat{\sigma}_\varepsilon^2$

$$\hat{\mu}_{iw} \quad = \quad \hat{\gamma}_{iw} \left[ \overline{y}_{iw} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_{iw})^{\top} \hat{\boldsymbol{\beta}}_w \right] + (1 - \hat{\gamma}_{iw}) \overline{\mathbf{X}}_i \hat{\boldsymbol{\beta}}_w,$$

where $\hat{\boldsymbol{\beta}}_w = \tilde{\boldsymbol{\beta}}_w(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2)$.

You & Rao (2002) showed that the estimators $\hat{\mu}_{iw}$ for $i = 1, \ldots, m$ automatically respect the benchmarking property as soon as the weights are calibrated on the known area sizes $N_i$ and when the intercept term is included in the nested error regression model (4.2). More precisely, we have

$$\sum_{i=1}^{m} N_i \hat{\mu}_{iw} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} y_{ij} + \left( \sum_{i=1}^{m} \sum_{j=1}^{N_i} \mathbf{x}_{ij} - \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} \mathbf{x}_{ij} \right)^{\top} \hat{\boldsymbol{\beta}}_w$$

where $\sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} y_{ij} + \left( \sum_{i=1}^{m} \sum_{j=1}^{N_i} \mathbf{x}_{ij} - \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} \mathbf{x}_{ij} \right)^{\top} \hat{\boldsymbol{\beta}}_w$ is the survey regression estimator of the overall total.

### 4.4.2 Reformulation of the pseudo-Eblup estimator

Like the Eblup estimator, the pseudo-Eblup estimator is clearly composite. It can be written as a weighted sum of two estimators: the weighted survey regression estimator $\hat{\mu}_{iw,dir}$ and the weighted regression synthetic estimator $\hat{\mu}_{iw,syn}$.

$$\begin{aligned}\hat{\mu}_{iw} &= \hat{\gamma}_{iw}\left[\overline{y}_{iw} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_{iw})^\top \hat{\boldsymbol{\beta}}_w\right] + (1-\hat{\gamma}_{iw})\overline{\mathbf{X}}_i^\top \hat{\boldsymbol{\beta}}_w, \\ &= \hat{\gamma}_{iw}\hat{\mu}_{iw,dir} + (1-\hat{\gamma}_{iw})\hat{\mu}_{iw,syn}. \end{aligned} \tag{4.21}$$

The two component estimators also are linear. In fact, from (4.19) we can write

$$\begin{aligned}\hat{\mu}_{iw,dir} &= \overline{y}_{iw} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_{iw})^\top \hat{\boldsymbol{\beta}}_w \\ &= \sum_{j=1}^{n_i} \frac{w_{ij}}{w_{i.}} y_{ij} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_{iw})^\top \left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}(\mathbf{x}_{ij} - \hat{\gamma}_{iw}\overline{\mathbf{x}}_{iw})^\top\right]^{-1} \\ &\quad \sum_{k=1}^{m}\sum_{j=1}^{n_k} w_{kj}(\mathbf{x}_{kj} - \hat{\gamma}_{kw}\overline{\mathbf{x}}_{kw})y_{kj} \\ &= \sum_{j=1}^{n_i} \frac{w_{ij}}{w_{i.}} y_{ij} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_{iw})^\top \mathbf{T}_w^{-1} \sum_{k=1}^{m}\sum_{j=1}^{n_k} w_{kj}(\mathbf{x}_{kj} - \hat{\gamma}_{kw}\overline{\mathbf{x}}_{kw})y_{kj} \\ &= \sum_{k=1}^{m}\sum_{j=1}^{n_k}\left[\frac{w_{kj}}{w_{k.}} y_{kj}\mathbb{1}_{\{k=i\}} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_{iw})^\top \mathbf{T}_w^{-1} w_{kj}(\mathbf{x}_{kj} - \hat{\gamma}_{kw}\overline{\mathbf{x}}_{kw})y_{kj}\right] \\ &= \sum_{k=1}^{m}\sum_{j=1}^{n_k}\left[\frac{w_{kj}}{w_{k.}}\mathbb{1}_{\{k=i\}} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_{iw})^\top \mathbf{T}_w^{-1} w_{kj}(\mathbf{x}_{kj} - \hat{\gamma}_{kw}\overline{\mathbf{x}}_{kw})\right] y_{kj} \\ &= \sum_{k=1}^{m}\sum_{j=1}^{n_k} w_{iw,dir}^{kj} y_{kj}, \end{aligned} \tag{4.22}$$

and

$$\begin{aligned}\hat{\mu}_{iw,syn} &= \overline{\mathbf{X}}_i^\top \left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}(\mathbf{x}_{ij} - \hat{\gamma}_{iw}\overline{\mathbf{x}}_{iw})^\top\right]^{-1} \sum_{k=1}^{m}\sum_{j=1}^{n_k} w_{kj}(\mathbf{x}_{kj} - \hat{\gamma}_{kw}\overline{\mathbf{x}}_{kw})y_{kj} \\ &= \sum_{k=1}^{m}\sum_{j=1}^{n_k}\left[\overline{\mathbf{X}}_i^\top \mathbf{T}_w^{-1} w_{kj}(\mathbf{x}_{kj} - \hat{\gamma}_{kw}\overline{\mathbf{x}}_{kw})\right] y_{kj} \\ &= \sum_{k=1}^{m}\sum_{j=1}^{n_k} w_{iw,syn}^{kj} y_{kj}. \end{aligned} \tag{4.23}$$

From combining (4.22) and (4.23), a version with weights of the pseudo-Eblup estimator is obtained as follows

$$\begin{aligned}\hat{\mu}_{iw} &= \hat{\gamma}_{iw}\sum_{k=1}^{m}\sum_{j=1}^{n_k} w_{iw,dir}^{kj} y_{kj} + (1-\hat{\gamma}_{iw})\sum_{k=1}^{m}\sum_{j=1}^{n_k} w_{iw,syn}^{kj} y_{kj} \\ &= \sum_{k=1}^{m}\sum_{j=1}^{n_k}\left[\hat{\gamma}_{iw} w_{i,dir}^{kj} + (1-\hat{\gamma}_{iw}) w_{iw,syn}^{kj}\right] y_{kj} \\ &= \sum_{k=1}^{m}\sum_{j=1}^{n_k} w_{iw}^{kj} y_{kj}. \end{aligned} \tag{4.24}$$

### 4.4.3 Calibration of the pseudo-Eblup estimator

Like the Eblup estimator and its component estimators, the pseudo-Eblup estimator and its component estimators are calibrated on the known auxiliary variables. The same as before, let us apply the pseudo-Eblup weights to the auxiliary variables $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^\top$ for $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$. We then have

$$\begin{aligned}\sum_{k=1}^{m}\sum_{j=1}^{n_k} w_{iw}^{kj}\mathbf{x}_{kj}^\top &= \sum_{k=1}^{m}\sum_{j=1}^{n_k}\left[\hat{\gamma}_{iw} w_{iw,dir}^{kj} + (1-\hat{\gamma}_{iw}) w_{iw,syn}^{kj}\right]\mathbf{x}_{kj}^\top \\ &= \hat{\gamma}_{iw}\left[\overline{\mathbf{x}}_{iw} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_{iw})^\top \mathbf{T}_w^{-1}\mathbf{T}_w\right] + (1-\hat{\gamma}_{iw})\overline{\mathbf{X}}_i^\top \mathbf{T}_w^{-1}\mathbf{T}_w \\ &= \hat{\gamma}_{iw}\overline{\mathbf{X}}_i + (1-\hat{\gamma}_{iw})\overline{\mathbf{X}}_i \\ &= \overline{\mathbf{X}}_i, \end{aligned} \tag{4.25}$$

provided that $\mathbf{T}_w$ is invertible.
Since calibrated, the pseudo-Eblup estimator also is then a precise estimator when the variable of interest is very correlated with the auxiliary variables.

## 4.5 A compromise weighting system

The equations (4.12), (4.13), (4.22) and (4.23) clearly show that the weighting forms of the Eblup and pseudo-Eblup estimators only depend on the tuning constants $\hat{\gamma}_i$ and $\hat{\gamma}_{iw}$, for $i = 1, \ldots, m$, which depend on the variable of interest.

We are now interested in using only a unique weighting system for a set of $Q$ variables of interest

$$\mathbf{Y}_i = (\mathbf{y}_i^1 \ldots \mathbf{y}_i^q \ldots \mathbf{y}_i^Q) = \begin{pmatrix} y_{i1}^1 & \ldots & y_{i1}^q & \ldots & y_{i1}^Q \\ \vdots & & \vdots & & \\ y_{ij}^1 & \ldots & y_{ij}^q & \ldots & y_{ij}^Q \\ \vdots & & \vdots & & \\ y_{in_i}^1 & \ldots & y_{in_i}^q & \ldots & y_{in_i}^Q \end{pmatrix},$$

for $i = 1, \ldots, m$, belonging to a specified theme, that is, the variables of interest are very correlated with each others.

For the set of variables of interest $\mathbf{Y}_i$ for $i = 1, \ldots, m$, the idea is to take common compromise tuning constants, $\hat{\Gamma}_i$, for the EBLUP estimator, and $\hat{\Gamma}_{iw}$, for the pseudo-EBLUP estimator. The common compromise tuning constants can be chosen as follows

$$\hat{\Gamma}_i = \frac{1}{Q} \sum_{q=1}^{Q} \hat{\gamma}_i^q, \text{ for } i = 1, \ldots, m \tag{4.26}$$

and

$$\hat{\Gamma}_{iw} = \frac{1}{Q} \sum_{q=1}^{Q} \hat{\gamma}_{iw}^q, \text{ for } i = 1, \ldots, m. \tag{4.27}$$

### 4.5.1 A compromise weighting system for the EBLUP estimator

From using the compromise tuning constant $\hat{\Gamma}_i$ for $i = 1, \ldots, m$, the component weights (4.12, 4.13) associated to the EBLUP estimator are re-written as follows, for $i = 1, \ldots, m$

$$\begin{aligned} w_{i,dir,mod}^{kj} &= \frac{1}{n_k} \mathbb{1}_{\{k=i\}} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_i)^\top \left[ \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbf{x}_{ij} (\mathbf{x}_{ij} - \hat{\Gamma}_i \overline{\mathbf{x}}_i)^\top \right]^{-1} (\mathbf{x}_{kj} - \hat{\Gamma}_k \overline{\mathbf{x}}_k), \\ w_{i,syn,mod}^{kj} &= \overline{\mathbf{X}}_i^\top \left[ \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbf{x}_{ij} (\mathbf{x}_{ij} - \hat{\Gamma}_i \overline{\mathbf{x}}_i)^\top \right]^{-1} (\mathbf{x}_{kj} - \hat{\Gamma}_k \overline{\mathbf{x}}_k). \end{aligned}$$

Then, the weights associated to the new EBLUP estimator are written as follows

$$w_{i,mod}^{kj} = \hat{\Gamma}_i w_{i,dir,mod}^{kj} + (1 - \hat{\Gamma}_i) w_{i,synth,mod}^{kj}. \tag{4.28}$$

For the $q$-th variable of the set $\mathbf{Y}_i$, the new EBLUP estimator which can be denoted by $\hat{\mu}_{i,Q}^q$, can be defined as

$$\hat{\mu}_{i,Q}^q = \sum_{k=1}^{m} \sum_{j=1}^{n_k} w_{i,mod}^{kj} y_{kj}^q. \tag{4.29}$$

Once again, the new EBLUP estimator and its component estimators are calibrated. In fact, for $i = 1, \ldots, m$,

$$\begin{aligned} \sum_{k=1}^{m} \sum_{j=1}^{n_k} w_{i,mod}^{kj} \mathbf{x}_{kj}^\top &= \hat{\Gamma}_i \sum_{k=1}^{m} \sum_{j=1}^{n_k} w_{i,dir,mod}^{kj} \mathbf{x}_{kj}^\top + (1 - \hat{\Gamma}_i) \sum_{k=1}^{m} \sum_{j=1}^{n_k} w_{i,synth,mod}^{kj} \mathbf{x}_{kj}^\top \\ &= \hat{\Gamma}_i \overline{\mathbf{X}}_i + (1 - \hat{\Gamma}_i) \overline{\mathbf{X}}_i \\ &= \overline{\mathbf{X}}_i, \end{aligned} \tag{4.30}$$

provided that the matrix $\sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbf{x}_{ij} (\mathbf{x}_{ij} - \hat{\Gamma}_i \overline{\mathbf{x}}_i)^\top$ is invertible.

### 4.5.2 A compromise weighting system for the pseudo-EBLUP estimator

The same as before, from using the compromise tuning constant $\hat{\Gamma}_{iw}$ for $i = 1, \ldots, m$, the component weights (4.22, 4.23) associated to the pseudo-EBLUP estimator are re-written as follows, for $i = 1, \ldots, m$

$$\begin{aligned} w^{kj}_{iw,dir,mod} &= \frac{w_{kj}}{w_{k.}}\mathbb{1}_{\{k=i\}} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_{iw})^\top \left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}(\mathbf{x}_{ij} - \hat{\Gamma}_{iw}\overline{\mathbf{x}}_{iw})^\top\right]^{-1} \\ & \quad w_{kj}(\mathbf{x}_{kj} - \hat{\Gamma}_{kw}\overline{\mathbf{x}}_{kw}), \\ w^{kj}_{iw,syn,mod} &= \overline{\mathbf{X}}_i^\top \left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}(\mathbf{x}_{ij} - \hat{\Gamma}_{iw}\overline{\mathbf{x}}_{iw})^\top\right]^{-1} w_{kj}(\mathbf{x}_{kj} - \hat{\Gamma}_{kw}\overline{\mathbf{x}}_{kw}). \end{aligned}$$

Then, the weights associated to the new pseudo-EBLUP estimator are written as follows

$$w^{kj}_{iw,mod} = \hat{\Gamma}_{iw}w^{kj}_{iw,dir,mod} + (1 - \hat{\Gamma}_{iw})w^{kj}_{iw,synth,mod}. \tag{4.31}$$

Once again, the new pseudo-EBLUP estimator and its component estimators are calibrated. In fact, for $i = 1, \ldots, m$,

$$\begin{aligned} \sum_{k=1}^{m}\sum_{j=1}^{n_k} w^{kj}_{iw,mod}\mathbf{x}^\top_{kj} &= \hat{\Gamma}_{iw}\sum_{k=1}^{m}\sum_{j=1}^{n_k} w^{kj}_{iw,dir,mod}\mathbf{x}^\top_{kj} + (1 - \hat{\Gamma}_{iw})\sum_{k=1}^{m}\sum_{j=1}^{n_k} w^{kj}_{iw,synth,mod}\mathbf{x}^\top_{kj} \\ &= \hat{\Gamma}_{iw}\overline{\mathbf{X}}_i + (1 - \hat{\Gamma}_{iw})\overline{\mathbf{X}}_i \\ &= \overline{\mathbf{X}}_i, \end{aligned} \tag{4.32}$$

provided that the matrix $\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}(\mathbf{x}_{ij} - \hat{\Gamma}_{iw}\overline{\mathbf{x}}_{iw})^\top$ is invertible.

The benchmarking property of the pseudo-EBLUP estimator is kept when using the unique tuning constant. For a given variable of interest $y^q_{ij}$ where for $j = 1, \ldots, n_i$, $k = 1, \ldots, m$, $i = 1, \ldots, m$ and $q = 1, \ldots, Q$ , we have

$$\begin{aligned} \sum_{i=1}^{m} N_i\hat{\mu}^q_{iw,Q} &= \sum_i N_i \sum_k\sum_j w^{kj}_{iw,mod}y^q_{kj} \\ &= \sum_i N_i \left[\hat{\Gamma}_{iw}\sum_k\sum_j w^{kj}_{iw,dir,mod}y^q_{kj} + (1 - \hat{\Gamma}_{iw})\sum_k\sum_j w^{kj}_{iw,synth,mod}y^q_{kj}\right] \\ &= \sum_i N_i \left[\hat{\Gamma}_{iw}\left(\overline{y}^q_{iw} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_{iw})^\top\left[\sum_i\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}(\mathbf{x}_{ij} - \hat{\Gamma}_{iw}\overline{\mathbf{x}}_{iw})^\top\right]^{-1}\right.\right. \\ &\quad \left.\left.\sum_k\sum_j w_{kj}(\mathbf{x}_{kj} - \hat{\Gamma}_{kw}\overline{\mathbf{x}}_{kw})y^q_{kj}\right) + (1 - \hat{\Gamma}_{iw})\overline{\mathbf{X}}_i^\top\left[\sum_i\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}\right. \right. \\ &\quad \left.\left.(\mathbf{x}_{ij} - \hat{\Gamma}_{iw}\overline{\mathbf{x}}_{iw})^\top\right]^{-1}\sum_k\sum_j w_{kj}(\mathbf{x}_{kj} - \hat{\Gamma}_{kw}\overline{\mathbf{x}}_{kw})y^q_{kj}\right] \\ &= \sum_i N_i \left[\overline{\mathbf{X}}_i^\top\hat{\boldsymbol{\beta}}^q_w + \hat{\Gamma}_{iw}(\overline{y}^q_{iw} - \overline{\mathbf{x}}^\top_{iw}\hat{\boldsymbol{\beta}}^q_w)\right], \end{aligned}$$

where $\hat{\boldsymbol{\beta}}_w^q = \left[\sum_{i=1}^m \sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}(\mathbf{x}_{ij} - \hat{\Gamma}_{iw}\bar{\mathbf{x}}_{iw})^\top\right]^{-1} \sum_{k=1}^m \sum_{j=1}^{n_k} w_{kj}(\mathbf{x}_{kj} - \hat{\Gamma}_{kw}\bar{\mathbf{x}}_{kw})y_{kj}^q$, and $\hat{\mu}_{iw,Q}^q$ denotes the new pseudo-EBLUP estimator.
From partially considering (4.20) corresponding to the intercept term, we have

$$\sum_{i=1}^m N_i\hat{\Gamma}_{iw}(\overline{y}_{iw}^q - \bar{\mathbf{x}}_{iw}^\top\hat{\boldsymbol{\beta}}_w^q) = \sum_{i=1}^{n_i}\sum_{j=1}^{n_i} w_{ij}(y_{ij}^q - \mathbf{x}_{ij}^\top\hat{\boldsymbol{\beta}}_w^q).$$

Finally, we obtain

$$\sum_{i=1}^m N_i\hat{\mu}_{iw,Q}^q = \sum_{i=1}^{n_i}\sum_{j=1}^{n_i} w_{ij}y_{ij}^q + \left(\sum_{i=1}^m\sum_{j=1}^{N_i}\mathbf{x}_{ij} - \sum_{i=1}^m\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}\right)^\top \hat{\boldsymbol{\beta}}_w^q$$

where $\sum_{i=1}^{n_i}\sum_{j=1}^{n_i} w_{ij}y_{ij}^q + \left(\sum_{i=1}^m\sum_{j=1}^{N_i}\mathbf{x}_{ij} - \sum_{i=1}^m\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}\right)^\top \hat{\boldsymbol{\beta}}_w^q$ is the survey regression estimator of the overall population for the $q$-th variable of interest.

## 4.6 Weighting system transferability

Let us now consider a new variable of interest $\mathbf{y}_i^p$ which belongs to the same specified theme as that of the set of $Q$ variables of interest $(\mathbf{y}_i^1 \dots \mathbf{y}_i^q \dots \mathbf{y}_i^Q)$, for $i = 1, \dots, m$. We are interested on the new EBLUP and pseudo-EBLUP estimators of $\overline{y}_i^p$ built with the compromise weighting system obtained from this set of $Q$ variables of interest. Let us denote $\hat{\mu}_{i,Q}^{p*}$ and $\hat{\mu}_{iw,Q}^{p*}$ these new estimators.

### 4.6.1 Weighting system transferability for the EBLUP estimator

From using the compromise weighting system (4.28), the new EBLUP estimator of $\overline{y}_i^p$, denoted $\hat{\mu}_{i,Q}^{p*}$, built with the compromise weighting system obtained from the set of $Q$ variables of interest, for $i = 1 \dots m$, can be defined as

$$\begin{aligned}\hat{\mu}_{i,Q}^{p*} &= \sum_{k=1}^m\sum_{j=1}^{n_k} w_{i,mod}^{kj}y_{kj}^p, \qquad (4.33)\\ &= \hat{\Gamma}_i\sum_{k=1}^m\sum_{j=1}^{n_k} w_{i,dir,mod}^{kj}y_{kj}^p + (1-\hat{\Gamma}_i)\sum_{k=1}^m\sum_{j=1}^{n_k} w_{i,synth,mod}^{kj}y_{kj}^p.\end{aligned}$$

As seen in (4.30), the obtained new estimator $\hat{\mu}_{i,Q}^{p*}$ is automatically calibrated on the auxiliary variables.

### 4.6.2 Weighting system transferability for the pseudo-EBLUP estimator

The same as before, the new EBLUP estimator of $\overline{y}_i^p$, denoted $\hat{\mu}_{iw,Q}^{p*}$, built with the compromise weighting system obtained from the set of $Q$ vari-

ables of interest, for $i = 1 \dots m$, can be defined as

$$\begin{aligned} \hat{\mu}^{p*}_{iw,Q} &= \sum_{k=1}^{m}\sum_{j=1}^{n_k} w^{kj}_{iw,mod} y^p_{kj}, \qquad (4.34) \\ &= \hat{\Gamma}_{iw}\sum_{k=1}^{m}\sum_{j=1}^{n_k} w^{kj}_{iw,dir,mod} y^p_{kj} + (1-\hat{\Gamma}_{iw})\sum_{k=1}^{m}\sum_{j=1}^{n_k} w^{kj}_{iw,synth,mod} y^p_{kj}. \end{aligned}$$

As seen in (4.32), the obtained new estimator $\hat{\mu}^{p*}_{iw,Q}$ is automatically calibrated on the auxiliary variables.

The benchmarking property is still kept for this new estimator. In fact, following the same demonstration as in Section 4.5.2, we have

$$\begin{aligned} \sum_{i=1}^{m} N_i \hat{\mu}^{p*}_{iw,Q} &= \sum_i N_i \sum_k \sum_j w^{kj}_{iw,mod} y^p_{kj} \\ &= \sum_i N_i \left[ \hat{\Gamma}_{iw} \sum_k \sum_j w^{kj}_{iw,dir,mod} y^p_{kj} + (1-\hat{\Gamma}_{iw}) \sum_k \sum_j w^{kj}_{iw,synth,mod} y^p_{kj} \right] \\ &= \sum_i N_i \left[ \hat{\Gamma}_{iw} \left( \overline{y}^p_{iw} + (\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_{iw})^\top \left[ \sum_i \sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}(\mathbf{x}_{ij} - \hat{\Gamma}_{iw}\overline{\mathbf{x}}_{iw})^\top \right]^{-1} \right. \right. \\ &\quad \left. \sum_k \sum_j w_{kj}(\mathbf{x}_{kj} - \hat{\Gamma}_{kw}\overline{\mathbf{x}}_{kw}) y^p_{kj} \right) + (1-\hat{\Gamma}_{iw}) \overline{\mathbf{X}}_i^\top \left[ \sum_i \sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij} \right. \\ &\quad \left. \left. (\mathbf{x}_{ij} - \hat{\Gamma}_{iw}\overline{\mathbf{x}}_{iw})^\top \right]^{-1} \sum_k \sum_j w_{kj}(\mathbf{x}_{kj} - \hat{\Gamma}_{kw}\overline{\mathbf{x}}_{kw}) y^p_{kj} \right] \\ &= \sum_i N_i \left[ \overline{\mathbf{X}}_i^\top \hat{\boldsymbol{\beta}}^{p*}_w + \hat{\Gamma}_{iw}(\overline{y}^p_{iw} - \overline{\mathbf{x}}_{iw}^\top \hat{\boldsymbol{\beta}}^{p*}_w) \right], \end{aligned}$$

where $\hat{\boldsymbol{\beta}}^{p*}_w = \left[\sum_{i=1}^m \sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}(\mathbf{x}_{ij} - \hat{\Gamma}_{iw}\overline{\mathbf{x}}_{iw})^\top\right]^{-1} \sum_{k=1}^m \sum_{j=1}^{n_k} w_{kj}(\mathbf{x}_{kj} - \hat{\Gamma}_{kw}\overline{\mathbf{x}}_{kw}) y^p_{kj}$. Considering the part of (4.20) coresponding to the intercept term,

$$\sum_{i=1}^{m} N_i \hat{\Gamma}_{iw}(\overline{y}^p_{iw} - \overline{\mathbf{x}}_{iw}^\top \hat{\boldsymbol{\beta}}^{p*}_w) = \sum_{i=1}^{n_i}\sum_{j=1}^{n_i} w_{ij}(y^p_{ij} - \mathbf{x}_{ij}^\top \hat{\boldsymbol{\beta}}^{p*}_w).$$

And we obtain

$$\sum_{i=1}^{m} N_i \hat{\mu}^{p*}_{iw,Q} = \sum_{i=1}^{n_i}\sum_{j=1}^{n_i} w_{ij} y^p_{ij} + \left( \sum_{i=1}^{m}\sum_{j=1}^{N_i} \mathbf{x}_{ij} - \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij} \right)^\top \hat{\boldsymbol{\beta}}^{p*}_w$$

which corresponds to the survey regression estimator of the overall population for the variable of interest $\mathbf{y}^p$.

## 4.7 Simulation study

From a design-based point of view, a set of simulation is run to compare the precision between the EBLUP and pseudo-EBLUP estimators, and the

EBLUP and pseudo-EBLUP estimators running with a compromise weighting system, and with a transferred weighting system. A population is generated from a model. But, when obtained, the population is considered as fixed.

### 4.7.1 Simulated data

#### Simulated population

A fixed population of $N = 2,000$ units, with $m = 20$ disjoint areas of sizes $(N_1, \ldots, N_m)$, is created from a linear mixed model given in (4.1):

$$\begin{aligned} y_{ij1} &= \mathbf{x}_{ij}^\top \boldsymbol{\beta} + v_i + \varepsilon_{ij}, \\ \mathbf{x}_{ij} &= \begin{pmatrix} 1 & x_{ij} \end{pmatrix}^\top \text{ with } x_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(20, \sigma_x^2 = 9), \\ \beta &= \begin{pmatrix} 12 & 0.4 \end{pmatrix}^\top, \\ v_i &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_v^2 = 4), \\ \varepsilon_{ij} &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2 = 1). \end{aligned}$$

Two other variables of interest are created as follows

$$\begin{aligned} y_{ij2} &= y_{ij1}^{0.2} \ln(\mathcal{N}(7, 0.0625)), \\ y_{ij3} &= y_{ij1}^{0.6} \ln(\mathcal{N}(17, 0.25)) \end{aligned}$$

in order to obtain high correlations. Table 4.1 shows the correlations between these created variables of interest. The figure 4.1 gives an overview of the generated population with the three variables of interest.

Table 4.1 – *Correlations between the variables of interest from the generated population.*

| | $\mathbf{y}^1$ | $\mathbf{y}^2$ | $\mathbf{y}^3$ |
|---|---|---|---|
| $\mathbf{y}^1$ | 1 | 0.8 | 0.7 |
| $\mathbf{y}^2$ | 0.8 | 1 | 0.9 |
| $\mathbf{y}^3$ | 0.7 | 0.9 | 1 |

#### Precision comparison

From the generated population, $B = 10,000$ samples of size $n = 200$ are drawn by a simple random sampling without replacement within areas. For each drawn sample, the area sizes are fixed (see Table 4.2). Within areas, the EBLUP estimators and the pseudo-EBLUP estimators (with their standard and modified forms) are computed to estimate the means $\overline{y}_i$, for $i = 1, \ldots, 20$. The relative root mean square error (%RRMSE) is used to quantify the performance of the estimators.

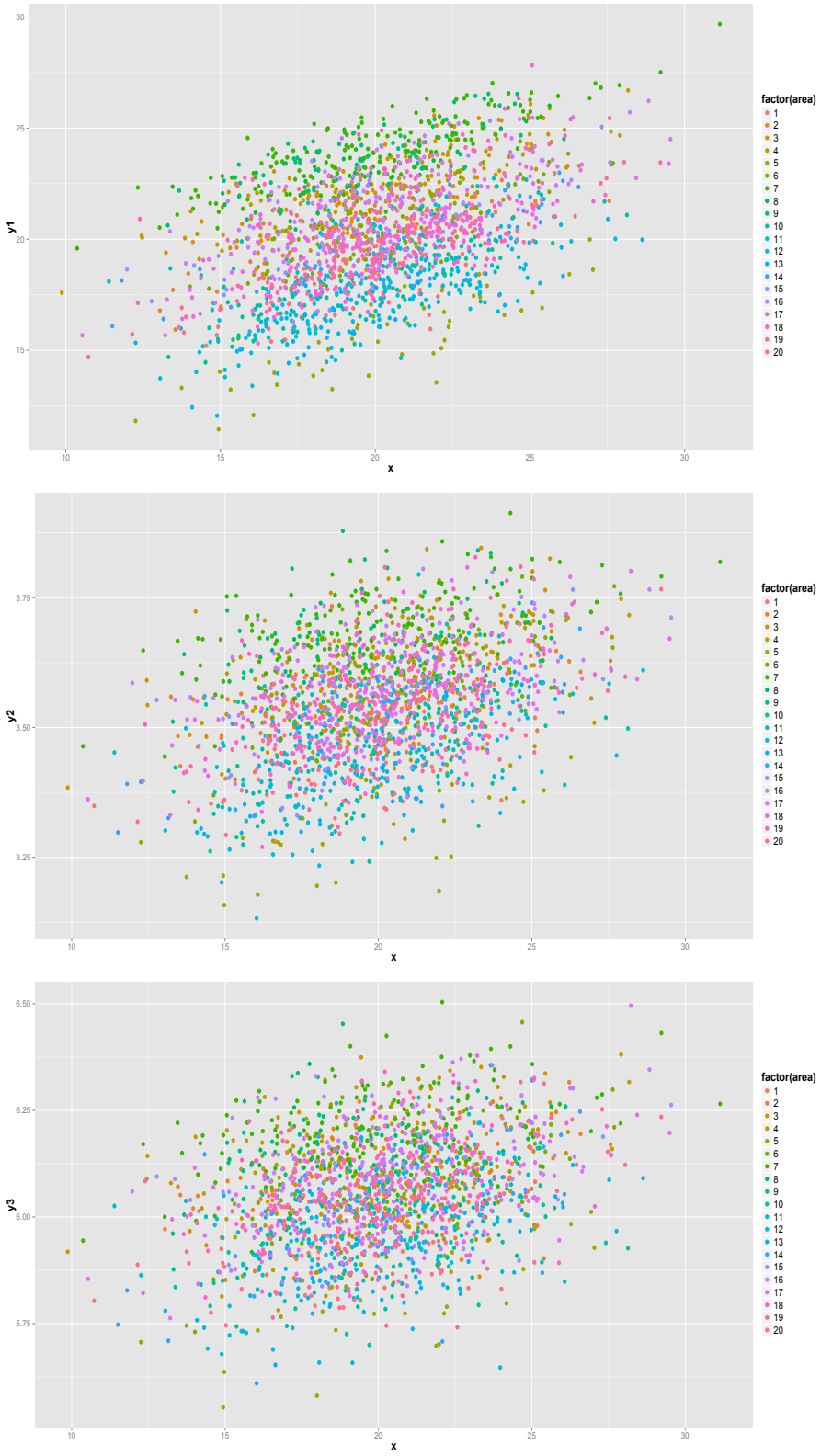


Figure 4.1 – *Generated population with the variables of interest* $\boldsymbol{y}^1$,$\boldsymbol{y}^2$ *and* $\boldsymbol{y}^3$.

For a given estimator $\hat{\bar{y}}_i$, the %RRMSE$_i$ is obtained as follows

$$\%\text{RRMSE}_i = 100 \times \frac{\sqrt{\text{MSE}_i}}{\bar{y}_i}, \tag{4.35}$$

where MSE$_i$ is the sum of the square of the bias and the variance

$$\text{MSE}_i = \left( \frac{1}{B} \sum_{b=1}^{B} \hat{\bar{y}}_i - \bar{y}_i \right)^2 + \frac{N-n}{N-1} \frac{1}{B-1} \sum_{b=1}^{B} \left( \hat{\bar{y}}_i - \frac{1}{B} \sum_{b=1}^{B} \bar{y}_i \right)^2 .$$

Table 4.2 – *Area sizes* $(N_1, \ldots, N_m)$ *in the population and fixed area sizes* $(n_1, \ldots, n_m)$ *in each drawn sample.*

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 2 | 11 | 7 | 16 | 3 | 15 | 11 | 6 | 5 | 21 |
| $N_i$ | 34 | 100 | 57 | 174 | 47 | 125 | 187 | 28 | 73 | 150 |
| $i$ | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| $n_i$ | 5 | 11 | 4 | 12 | 1 | 11 | 26 | 18 | 4 | 11 |
| $N_i$ | 44 | 182 | 35 | 78 | 49 | 124 | 168 | 116 | 65 | 164 |

### Simulation results

Rao (2003) shows that under a model-based inference, the EBLUP estimator is a quite more efficient than the pseudo-EBLUP estimator. The simulation results given by Tables 4.3, 4.4 and 4.5, and Tables 4.6, 4.7 and 4.8 show that, even under a design-based inference, the EBLUP estimator seems to be more efficient than the pseudo-EBLUP. This loss of efficiency can partly be the result of the constraint of benchmarking.

**Compromise weighting system** Considering the set of the three variables of interest $\mathbf{y}^1, \mathbf{y}^2$ and $\mathbf{y}^3$, we compare the precision of the standard forms of the EBLUP and pseudo-EBLUP estimators with that of the modified forms of the EBLUP and pseudo-EBLUP estimators using a compromise weighting system from the three variables.

Tables 4.3, 4.4 and 4.5 show that the precision of the obtained estimators using the compromise weighting system is very close to that of the standard forms of the EBLUP and pseudo-EBLUP estimators. However, at the level of some areas, especially at the level of the extremely small areas (for instance the areas 1,5 and 15), when the modified forms of the EBLUP and the pseudo-EBLUP estimators behave worse than the standard forms for a variable of interest, the precision of the modified forms for the other variables seems to improve. It seems to appear that there is an exchange of precision between the variables. When a variable looses some of its precision, another one catches this precision in order to improve its own.

Table 4.3 – *Computed %RRMSE of the standard forms of the* Eblup *and pseudo-*Eblup *estimators, and of the modified forms of the* Eblup *and pseudo-*Eblup *estimators using a compromise weighting system, for the variable* $\boldsymbol{y}^1$.

| area | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Eblup | standard | 3.34 | 1.44 | 1.31 | 1.06 | 3.92 | 1.10 | 1.06 | 1.38 |
| | modified | 5.15 | 1.37 | 1.61 | 1.08 | 6.96 | 1.06 | 1.43 | 1.85 |
| Pseudo-Eblup | standard | 3.35 | 1.44 | 1.31 | 1.06 | 3.92 | 1.10 | 1.07 | 1.38 |
| | modified | 5.11 | 1.37 | 1.62 | 1.09 | 6.93 | 1.06 | 1.45 | 1.86 |
| area | | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Eblup | standard | 1.80 | 1.03 | 2.02 | 1.81 | 2.28 | 1.36 | 4.75 | 1.24 |
| | modified | 2.55 | 1.10 | 1.83 | 2.13 | 3.06 | 1.35 | 3.51 | 1.30 |
| Pseudo-Eblup | standard | 1.81 | 1.03 | 2.02 | 1.81 | 2.27 | 1.36 | 4.84 | 1.24 |
| | modified | 2.58 | 1.10 | 1.83 | 2.12 | 3.03 | 1.35 | 3.65 | 1.31 |
| area | | 17 | 18 | 19 | 20 | | | | |
| Eblup | standard | 0.78 | 1.10 | 2.09 | 1.23 | | | | |
| | modified | 0.76 | 1.09 | 2.51 | 1.22 | | | | |
| Pseudo-Eblup | standard | 0.78 | 1.10 | 2.09 | 1.23 | | | | |
| | modified | 0.76 | 1.09 | 2.55 | 1.22 | | | | |

Table 4.4 – *Computed %RRMSE of the standard forms of the* Eblup *and pseudo-*Eblup *estimators, and of the modified forms of the* Eblup *and pseudo-*Eblup *estimators using a compromise weighting system, for the variable* $\boldsymbol{y}^2$.

| area | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Eblup | standard | 1.27 | 0.52 | 0.68 | 0.46 | 1.63 | 0.45 | 0.55 | 0.65 |
| | modified | 1.25 | 0.52 | 0.68 | 0.46 | 1.59 | 0.45 | 0.55 | 0.64 |
| Pseudo-Eblup | standard | 1.26 | 0.52 | 0.68 | 0.46 | 1.62 | 0.45 | 0.56 | 0.65 |
| | modified | 1.24 | 0.52 | 0.68 | 0.46 | 1.58 | 0.45 | 0.56 | 0.65 |
| area | | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Eblup | standard | 0.95 | 0.40 | 0.71 | 0.65 | 0.79 | 0.53 | 1.32 | 0.53 |
| | modified | 0.94 | 0.40 | 0.71 | 0.65 | 0.79 | 0.53 | 1.40 | 0.53 |
| Pseudo-Eblup | standard | 0.96 | 0.40 | 0.71 | 0.65 | 0.78 | 0.53 | 1.40 | 0.53 |
| | modified | 0.95 | 0.40 | 0.72 | 0.65 | 0.78 | 0.53 | 1.46 | 0.53 |
| area | | 17 | 18 | 19 | 20 | | | | |
| Eblup | standard | 0.31 | 0.43 | 0.94 | 0.56 | | | | |
| | modified | 0.31 | 0.43 | 0.93 | 0.55 | | | | |
| Pseudo-Eblup | standard | 0.31 | 0.43 | 0.95 | 0.56 | | | | |
| | modified | 0.31 | 0.43 | 0.94 | 0.56 | | | | |

Whereas the different obtained precisions for the variable $\mathbf{y}^2$ seem to be very close to each other, the precision exchange seem to be more significant between the variables $\mathbf{y}^1$ and $\mathbf{y}^3$. The Figure 4.2 gives an overview of the precision of the four considered estimators.

**Transferred weighting system** Considering the set of the three variables of interest $\mathbf{y}^1$, $\mathbf{y}^2$ and $\mathbf{y}^3$, we compare the precision of the standard forms of the Eblup and pseudo-Eblup estimators with that of the modified forms of the Eblup and pseudo-Eblup estimators using successive compromise

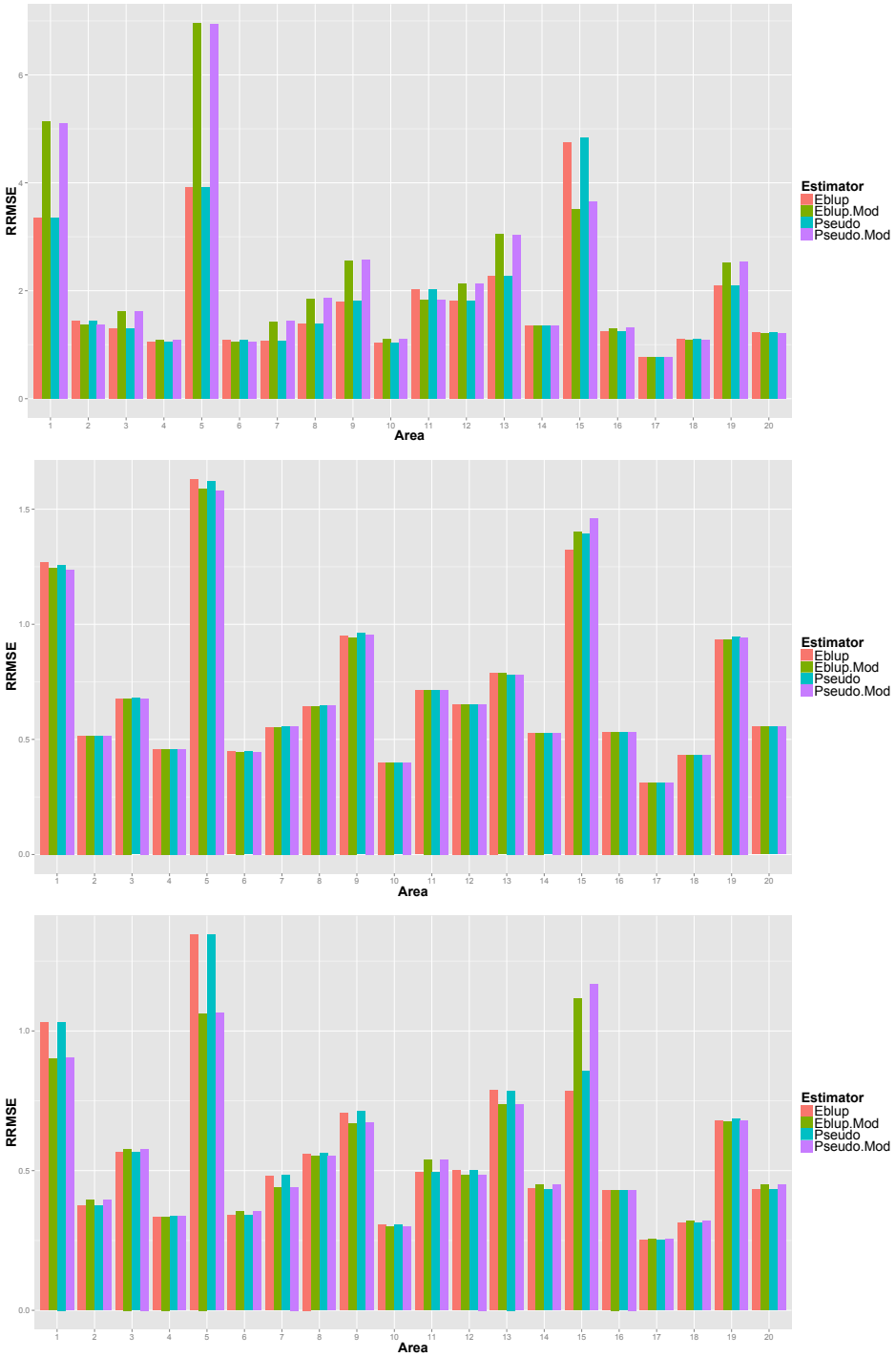


Figure 4.2 – *Barplots of the computed %RRMSE of the standard forms of the* Eblup *and pseudo-*Eblup *estimators, and of the modified forms of the* Eblup *and pseudo-*Eblup *estimators using a compromise weighting system, for the variables* $\boldsymbol{y}^1, \boldsymbol{y}^2$ *and* $\boldsymbol{y}^3$.

Table 4.5 – *Computed %RRMSE of the standard forms of the* Eblup *and pseudo-*Eblup *estimators, and of the modified forms of the* Eblup *and pseudo-*Eblup *estimators using a compromise weighting system, for the variable* $\boldsymbol{y}^3$.

| area | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Eblup | standard | 1.03 | 0.37 | 0.56 | 0.33 | 1.34 | 0.34 | 0.48 | 0.56 |
| | modified | 0.90 | 0.39 | 0.58 | 0.33 | 1.06 | 0.35 | 0.44 | 0.55 |
| Pseudo-Eblup | standard | 1.03 | 0.37 | 0.57 | 0.34 | 1.35 | 0.34 | 0.48 | 0.56 |
| | modified | 0.90 | 0.39 | 0.58 | 0.34 | 1.06 | 0.35 | 0.44 | 0.55 |
| area | | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Eblup | standard | 0.71 | 0.31 | 0.49 | 0.50 | 0.79 | 0.43 | 0.78 | 0.43 |
| | modified | 0.67 | 0.30 | 0.54 | 0.48 | 0.74 | 0.45 | 1.11 | 0.43 |
| Pseudo-Eblup | standard | 0.71 | 0.31 | 0.49 | 0.50 | 0.79 | 0.43 | 0.86 | 0.43 |
| | modified | 0.67 | 0.30 | 0.54 | 0.49 | 0.74 | 0.45 | 1.17 | 0.43 |
| area | | 17 | 18 | 19 | 20 | | | | |
| Eblup | standard | 0.25 | 0.31 | 0.68 | 0.43 | | | | |
| | modified | 0.26 | 0.32 | 0.67 | 0.45 | | | | |
| Pseudo-Eblup | standard | 0.25 | 0.31 | 0.69 | 0.43 | | | | |
| | modified | 0.26 | 0.32 | 0.68 | 0.45 | | | | |

weighting system from variables $(\mathbf{y}^2, \mathbf{y}^3)$, $(\mathbf{y}^1, \mathbf{y}^3)$, $(\mathbf{y}^1, \mathbf{y}^2)$ for estimating the means of $\mathbf{y}^1, \mathbf{y}^2$ and $\mathbf{y}^3$.

Table 4.6 – *Computed %RRMSE of the standard forms of the* Eblup *and pseudo-*Eblup *estimators, and of the modified forms of the* Eblup *and pseudo-*Eblup *estimators using the compromise weighting system from variables* $(\boldsymbol{y}^2, \boldsymbol{y}^3)$, *for the variable* $\boldsymbol{y}^1$.

| area | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Eblup | standard | 3.34 | 1.44 | 1.31 | 1.06 | 3.92 | 1.10 | 1.06 | 1.38 |
| | modified | 6.44 | 1.33 | 1.88 | 1.12 | 8.93 | 1.04 | 1.73 | 2.22 |
| Pseudo-Eblup | standard | 3.35 | 1.44 | 1.31 | 1.06 | 3.92 | 1.10 | 1.07 | 1.38 |
| | modified | 6.38 | 1.33 | 1.90 | 1.13 | 8.89 | 1.04 | 1.75 | 2.24 |
| area | | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Eblup | standard | 1.80 | 1.03 | 2.02 | 1.81 | 2.28 | 1.36 | 4.75 | 1.24 |
| | modified | 3.13 | 1.18 | 1.74 | 2.43 | 3.72 | 1.38 | 2.93 | 1.39 |
| Pseudo-Eblup | standard | 1.81 | 1.03 | 2.02 | 1.81 | 2.27 | 1.36 | 4.84 | 1.24 |
| | modified | 3.17 | 1.17 | 1.74 | 2.42 | 3.67 | 1.37 | 3.11 | 1.40 |
| area | | 17 | 18 | 19 | 20 | | | | |
| Eblup | standard | 0.78 | 1.10 | 2.09 | 1.23 | | | | |
| | modified | 0.75 | 1.10 | 2.92 | 1.24 | | | | |
| Pseudo-Eblup | standard | 0.78 | 1.10 | 2.09 | 1.23 | | | | |
| | modified | 0.75 | 1.10 | 2.96 | 1.23 | | | | |

The same as before, Tables 4.6, 4.7 and 4.8 show that the precision of the obtained estimators using the transferred weighting system is close to that of the standard forms of the Eblup and pseudo-Eblup estimators. Similarly as the results from the compromise weighting system, the obtained precisions for the variable $\mathbf{y}^2$ are the closest, compared to those for the

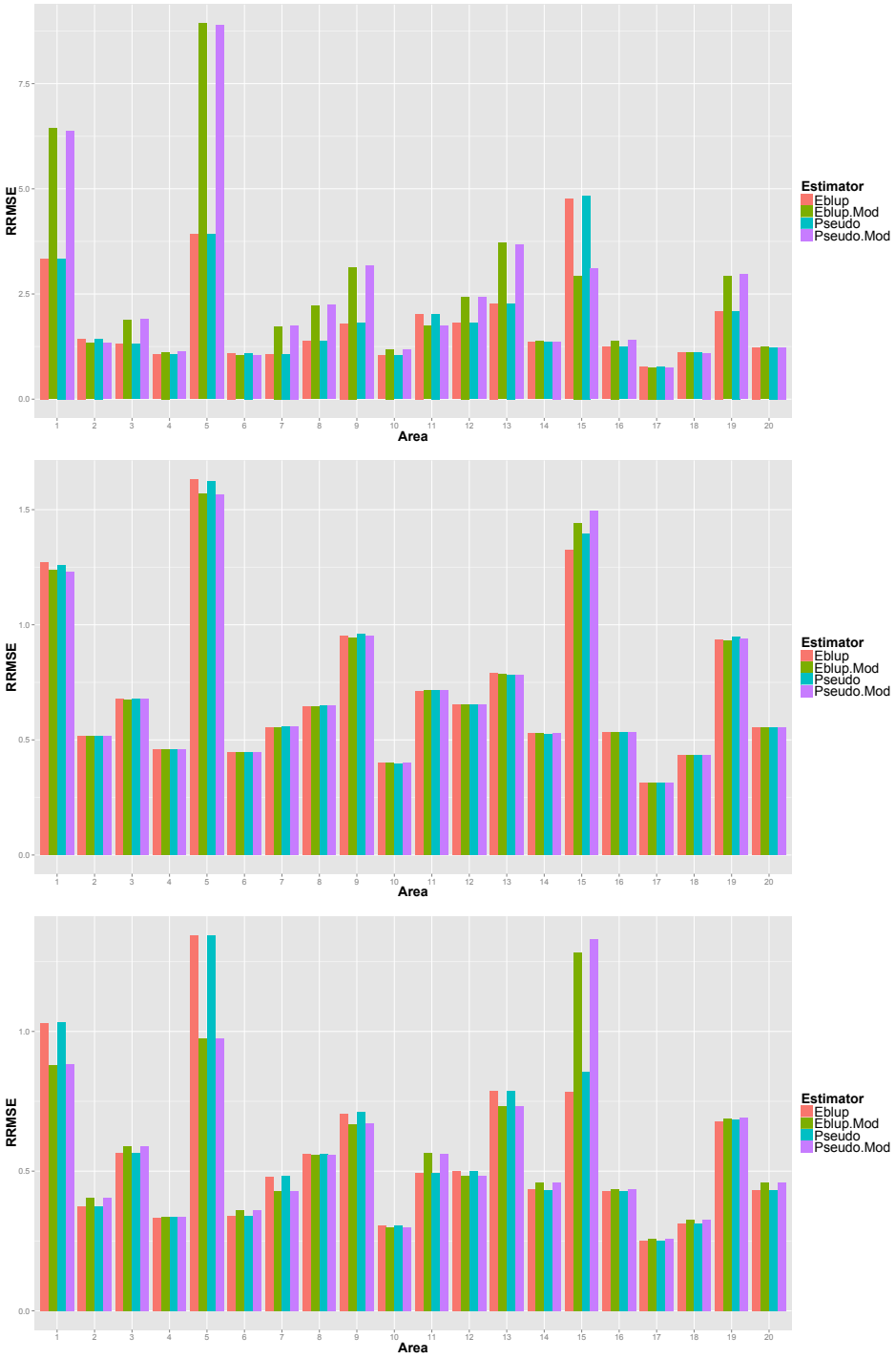


Figure 4.3 – *Barplots of the computed %RRMSE of the standard forms of the* Eblup *and pseudo-*Eblup *estimators, and of the modified forms of the* Eblup *and pseudo-*Eblup *estimators using the compromise weighting systems from variables* $(\boldsymbol{y}^2, \boldsymbol{y}^3)$, $(\boldsymbol{y}^1, \boldsymbol{y}^3)$ *and* $(\boldsymbol{y}^1, \boldsymbol{y}^2)$, *for the variables* $\boldsymbol{y}^1, \boldsymbol{y}^2$ *and* $\boldsymbol{y}^3$.

Table 4.7 – *Computed %RRMSE of the standard forms of the* Eblup *and pseudo-*Eblup *estimators, and of the modified forms of the* Eblup *and pseudo-*Eblup *estimators using the compromise weighting system from variables* $(\boldsymbol{y}^1, \boldsymbol{y}^3)$*, for the variable* $\boldsymbol{y}^2$.

| area | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Eblup | standard | 1.27 | 0.52 | 0.68 | 0.46 | 1.63 | 0.45 | 0.55 | 0.65 |
| | modified | 1.24 | 0.52 | 0.68 | 0.46 | 1.57 | 0.45 | 0.55 | 0.64 |
| Pseudo-Eblup | standard | 1.26 | 0.52 | 0.68 | 0.46 | 1.62 | 0.45 | 0.56 | 0.65 |
| | modified | 1.23 | 0.52 | 0.68 | 0.46 | 1.56 | 0.45 | 0.56 | 0.65 |
| area | | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Eblup | standard | 0.95 | 0.40 | 0.71 | 0.65 | 0.79 | 0.53 | 1.32 | 0.53 |
| | modified | 0.94 | 0.40 | 0.72 | 0.65 | 0.79 | 0.53 | 1.44 | 0.53 |
| Pseudo-Eblup | standard | 0.96 | 0.40 | 0.71 | 0.65 | 0.78 | 0.53 | 1.40 | 0.53 |
| | modified | 0.95 | 0.40 | 0.72 | 0.65 | 0.78 | 0.53 | 1.50 | 0.53 |
| area | | 17 | 18 | 19 | 20 | | | | |
| Eblup | standard | 0.31 | 0.43 | 0.94 | 0.56 | | | | |
| | modified | 0.31 | 0.43 | 0.93 | 0.55 | | | | |
| Pseudo-Eblup | standard | 0.31 | 0.43 | 0.95 | 0.56 | | | | |
| | modified | 0.31 | 0.43 | 0.94 | 0.56 | | | | |

Table 4.8 – *Computed %RRMSE of the standard forms of the* Eblup *and pseudo-*Eblup *estimators, and of the modified forms of the* Eblup *and pseudo-*Eblup *estimators using the compromise weighting system from variables* $(\boldsymbol{y}^1, \boldsymbol{y}^2)$*, for the variable* $\boldsymbol{y}^3$.

| area | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Eblup | standard | 1.03 | 0.37 | 0.56 | 0.33 | 1.34 | 0.34 | 0.48 | 0.56 |
| | modified | 0.88 | 0.40 | 0.59 | 0.34 | 0.97 | 0.36 | 0.43 | 0.56 |
| Pseudo-Eblup | standard | 1.03 | 0.37 | 0.57 | 0.34 | 1.35 | 0.34 | 0.48 | 0.56 |
| | modified | 0.88 | 0.40 | 0.59 | 0.34 | 0.97 | 0.36 | 0.43 | 0.56 |
| area | | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Eblup | standard | 0.71 | 0.31 | 0.49 | 0.50 | 0.79 | 0.43 | 0.78 | 0.43 |
| | modified | 0.67 | 0.30 | 0.56 | 0.48 | 0.73 | 0.46 | 1.28 | 0.44 |
| Pseudo-Eblup | standard | 0.71 | 0.31 | 0.49 | 0.50 | 0.79 | 0.43 | 0.86 | 0.43 |
| | modified | 0.67 | 0.30 | 0.56 | 0.48 | 0.73 | 0.46 | 1.33 | 0.44 |
| area | | 17 | 18 | 19 | 20 | | | | |
| Eblup | standard | 0.25 | 0.31 | 0.68 | 0.43 | | | | |
| | modified | 0.26 | 0.33 | 0.69 | 0.46 | | | | |
| Pseudo-Eblup | standard | 0.25 | 0.31 | 0.69 | 0.43 | | | | |
| | modified | 0.26 | 0.33 | 0.69 | 0.46 | | | | |

variables $\mathbf{y}^1$ and $\mathbf{y}^3$. It can be explained by the fact that the variable $\mathbf{y}^2$ is as highly correlated with the variable $\mathbf{y}^1$ as with the variable $\mathbf{y}^3$. The correlation between the variables $\mathbf{y}^1$ and $\mathbf{y}^3$ is lower. The precision exchange seem to be significant between the variables $\mathbf{y}^1$ and $\mathbf{y}^3$. When the precision of the modified forms of the Eblup and the pseudo-Eblup estimators for the variable $\mathbf{y}^1$ decreases, the precision of the modified forms of the Eblup and the pseudo-Eblup estimators for the variable $\mathbf{y}^3$ increases, and vice versa. The Figure 4.3 gives an overview of the precision of the four considered estimators.

### 4.7.2 Application : data with county crop areas

We compare the EBLUP and the pseudo-EBLUP estimators, with their respective modified forms, applied to a real data given by Battese et al. (1988) and taken up by You & Rao (2002); Rao (2003). We want to estimate the means of hectares of corn, and of soybeans, per segment for $m = 12$ counties in north-central Iowa. The county is divided into area segments and the areas under corn and under soybeans are located in the area segments. A sample $s$ of $n = 36$ segments is used and simple random sampling within areas is assumed. The population is assumed to follow the linear mixed model given in (4.1). The functions of interest are the numbers of hectares of corn and of soybeans per segment per county. The auxiliary information is the numbers of pixels seen as corn and as soybeans.

As simple random sampling within areas is assumed, the tuning constants $\gamma_i$, $i = 1 \dots m$, of the EBLUP estimator coincide with the tuning constants $\gamma_{iw}$ of the pseudo-EBLUP. In fact, the survey weight for the sample unit $j$ in area $i$ is given by $w_{ij} = N_i/n_i$. Then, we have from (4.5) and (4.18) that for $i = 1 \dots m$, $\delta_{iw} = \sum_{j=1}^{n_i} (w_{ij}/w_{i.})^2 = \sum_{j=1}^{n_i} 1/n_i^2 = 1/n_i$, thus $\gamma_i = \gamma_{iw}$.

Coefficients of variation are used to compare the precision of the obtained estimators. The coefficients of variation are obtained by a resampling procedure of 10,000 iterations (see Algorithm 4.1).

**Algorithm 4.1:** Resampling procedure for estimating the coefficients of variation by bootstrap

1 Consider $B = 10,000$ the number of iterations;
2 Consider $n = 36$ the sample size;
3 **for** *each iteration* $b$ **do**
4 Draw a sample $s^b$ of size $n$ by a simple random sampling with replacement within areas from the sample $s$ (the area sample sizes are fixed and are the same as those in $s$);
5 Compute the EBLUP and pseudo-EBLUP estimators, with their respective modified forms from the sample $s^b$
6 Compute the variances of the $B$ obtained estimators;
7 Compute the EBLUP and pseudo-EBLUP estimators, with their respective modified forms from the sample $s$ with considering the finite population correction factor;
8 Compute the coefficients of variation;

Tables 4.9 and 4.10 compare the obtained precisions of the EBLUP and the pseudo-EBLUP estimates of the hectares of corn. Tables 4.11 and 4.12 com-

Table 4.9 – EBLUP *estimates of hectares of corn with bootstrap coefficients of variation estimates.*

| | | EBLUP | | | | | |
|---|---|---|---|---|---|---|---|
| | | Standard form | | Weighting system | | | |
| | | | | Compromise | | Transfer | |
| County | $n_i$ | Estimate | c.v. (%) | Estimate | c.v. (%) | Estimate | c.v. (%) |
| | | Corn | | | | | |
| Cerro Gordo | 1 | 122.2 | 1.2 | 122.0 | 1.3 | 121.8 | 1.4 |
| Hamilton | 1 | 126.2 | 1.3 | 126.5 | 1.3 | 126.7 | 1.4 |
| Worth | 1 | 106.7 | 0.8 | 105.8 | 0.8 | 104.8 | 0.8 |
| Humboldt | 2 | 108.4 | 7.8 | 107.9 | 8.2 | 107.3 | 8.7 |
| Franklin | 3 | 144.3 | 3.4 | 144.8 | 3.5 | 145.3 | 3.6 |
| Pocahontas | 3 | 112.1 | 4.2 | 112.5 | 4.3 | 112.9 | 4.4 |
| Winnebago | 3 | 112.8 | 4.9 | 112.4 | 5.1 | 112.0 | 5.3 |
| Wright | 3 | 122.0 | 3.5 | 122.0 | 3.6 | 122.0 | 3.7 |
| Webster | 4 | 115.3 | 2.4 | 115.6 | 2.5 | 116.0 | 2.5 |
| Hancock | 5 | 124.4 | 3.1 | 124.4 | 3.2 | 124.4 | 3.3 |
| Kossuth | 5 | 106.9 | 1.9 | 106.5 | 1.9 | 106.1 | 2.0 |
| Hardin | 5 | 143.0 | 2.8 | 143.3 | 2.8 | 143.6 | 2.9 |

Source: LANDSAT data from Table 1 in Battese et al. (1988, p. 29).

Table 4.10 – *Pseudo*-EBLUP *estimates of hectares of corn with bootstrap coefficients of variation estimates.*

| | | Pseudo-EBLUP | | | | | |
|---|---|---|---|---|---|---|---|
| | | Standard form | | Weighting system | | | |
| | | | | Compromise | | Transfer | |
| County | $n_i$ | Estimate | c.v. (%) | Estimate | c.v. (%) | Estimate | c.v. (%) |
| | | Corn | | | | | |
| Cerro Gordo | 1 | 120.5 | 1.5 | 120.2 | 1.6 | 120.0 | 1.8 |
| Hamilton | 1 | 125.3 | 1.4 | 125.6 | 1.4 | 125.8 | 1.5 |
| Worth | 1 | 106.3 | 1.1 | 105.5 | 1.1 | 104.6 | 1.1 |
| Humboldt | 2 | 107.3 | 8.1 | 106.8 | 8.6 | 106.2 | 9.0 |
| Franklin | 3 | 143.8 | 3.3 | 144.3 | 3.4 | 144.8 | 3.5 |
| Pocahontas | 3 | 111.5 | 4.0 | 111.9 | 4.1 | 112.4 | 4.2 |
| Winnebago | 3 | 112.1 | 5.0 | 111.7 | 5.1 | 111.3 | 5.3 |
| Wright | 3 | 121.3 | 3.8 | 121.3 | 3.9 | 121.3 | 4.0 |
| Webster | 4 | 115.1 | 2.5 | 115.4 | 2.5 | 115.8 | 2.6 |
| Hancock | 5 | 124.5 | 3.1 | 124.5 | 3.2 | 124.6 | 3.3 |
| Kossuth | 5 | 106.6 | 2.3 | 106.2 | 2.3 | 105.8 | 2.4 |
| Hardin | 5 | 143.5 | 2.9 | 143.9 | 2.9 | 144.2 | 3.0 |

Source: LANDSAT data from Table 1 in Battese et al. (1988, p. 29).

pare the obtained precisions of the EBLUP and the pseudo-EBLUP estimates of the hectares of soybeans. The compromise weighting system is built from these two variables. The transferred weighting system, used for the estimation of the hectares of corn, comes from the variable *hectares of corn*, and vice versa, the transferred weighting system, used for the estimation of the hectares of corn, comes from the variable *hectares of soybeans*. Both

Table 4.11 – EBLUP *estimates of hectares of soybeans with bootstrap coefficients of variation estimates.*

| | | EBLUP | | | | | |
|---|---|---|---|---|---|---|---|
| | | Standard form | | Weighting system | | | |
| | | | | Compromise | | Transfer | |
| County | $n_i$ | Estimate | c.v. (%) | Estimate | c.v. (%) | Estimate | c.v. (%) |
| | | Soybeans | | | | | |
| Cerro Gordo | 1 | 78.5 | 2.6 | 79.2 | 2.3 | 79.8 | 2.2 |
| Hamilton | 1 | 94.4 | 2.4 | 94.1 | 2.3 | 93.7 | 2.2 |
| Worth | 1 | 87.4 | 0.9 | 87.9 | 0.9 | 88.3 | 0.9 |
| Humboldt | 2 | 81.1 | 4.9 | 82.2 | 4.5 | 83.3 | 4.1 |
| Franklin | 3 | 66.2 | 4.5 | 67.0 | 4.3 | 67.7 | 4.2 |
| Pocahontas | 3 | 113.7 | 7.6 | 113.7 | 7.3 | 113.7 | 6.9 |
| Winnebago | 3 | 97.8 | 8.3 | 97.3 | 8.1 | 96.9 | 7.9 |
| Wright | 3 | 112.3 | 5.5 | 112.0 | 5.3 | 111.7 | 5.1 |
| Webster | 4 | 109.8 | 3.0 | 109.9 | 2.9 | 110.1 | 2.8 |
| Hancock | 5 | 100.7 | 5.3 | 100.4 | 5.3 | 100.2 | 5.2 |
| Kossuth | 5 | 119.0 | 3.6 | 118.4 | 3.6 | 117.9 | 3.5 |
| Hardin | 5 | 75.2 | 6.4 | 75.2 | 6.3 | 75.3 | 6.2 |

Source: LANDSAT data from Table 1 in Battese et al. (1988, p. 29).

Table 4.12 – *Pseudo-*EBLUP *estimates of hectares of soybeans with bootstrap coefficients of variation estimates.*

| | | Pseudo-EBLUP | | | | | |
|---|---|---|---|---|---|---|---|
| | | Standard form | | Weighting system | | | |
| | | | | Compromise | | Transfer | |
| County | $n_i$ | Estimate | c.v. (%) | Estimate | c.v. (%) | Estimate | c.v. (%) |
| | | Soybeans | | | | | |
| Cerro Gordo | 1 | 80.3 | 3.2 | 80.7 | 2.9 | 81.1 | 2.6 |
| Hamilton | 1 | 92.3 | 2.5 | 92.0 | 2.4 | 91.7 | 2.3 |
| Worth | 1 | 85.9 | 1.3 | 86.4 | 1.3 | 86.8 | 1.3 |
| Humboldt | 2 | 83.8 | 6.3 | 84.7 | 5.8 | 85.6 | 5.3 |
| Franklin | 3 | 66.5 | 4.6 | 67.1 | 4.4 | 67.8 | 4.2 |
| Pocahontas | 3 | 113.0 | 8.6 | 113.0 | 8.2 | 113.1 | 7.9 |
| Winnebago | 3 | 97.8 | 7.6 | 97.3 | 7.4 | 96.8 | 7.2 |
| Wright | 3 | 113.5 | 6.1 | 113.2 | 5.9 | 112.8 | 5.6 |
| Webster | 4 | 109.7 | 3.0 | 109.9 | 2.9 | 110.0 | 2.8 |
| Hancock | 5 | 99.2 | 4.8 | 99.0 | 4.8 | 98.7 | 4.7 |
| Kossuth | 5 | 119.3 | 3.5 | 118.7 | 3.4 | 118.1 | 3.4 |
| Hardin | 5 | 72.8 | 5.3 | 72.8 | 5.2 | 72.9 | 5.1 |

Source: LANDSAT data from Table 1 in Battese et al. (1988, p. 29).

for the hectares of corn and of soybeans, the pseudo-EBLUP estimates lose efficiency compared to the EBLUP estimates.

When using the compromise and the transferred weighting systems, the precisions of the EBLUP and the pseudo-EBLUP estimates of the hectares of corn decrease. Conversely, the precisions of the EBLUP and the pseudo-

EBLUP estimates of the hectares of soybeans increase. It can confirm the existence of an exchange of precision between the two considered variables.

## 4.8 Concluding remarks

The set of simulations shows that the proposed modified forms of the EBLUP and pseudo-EBLUP estimators have efficiency close to that of the classical forms. Overall, the loss of precision from the classical forms to the modified forms is small. The proposed modified forms draw their advantage on the practicality from using weights.

Since the proposed modified forms of the EBLUP and pseudo-EBLUP estimators clearly have complex forms, the calculation of the variances also is complex. Then, as it is done in the set of simulations, resampling methods can be used.

# 5 Estimation régionale de la mobilité locale en France à partir de l'Enquête Nationale sur les Transports et les Déplacements 2007-2008

**Résumé** Ce chapitre présente les différents résultats obtenus de l'estimation régionale de la mobilité en France. Les Sections 5.2 et 5.3 présentent l'Enquête Nationale sur les Transports et les Déplacements 2007-2008. Ces deux sections reprennent la présentation de l'Enquête Nationale sur les Transports et les Déplacements 2007-2008 donnée dans les notes méthodologiques de Roux & Armoogum (2008, 2010) et de Randrianasolo et al. (2010). La Section 5.4 présente la pondération de l'Enquête Nationale sur les Transports et les Déplacements 2007-2008 effectuée par Roux & Armoogum (2010); Armoogum & Roux (2012). L'objectif principal de ce chapitre est présenté dans la Section 5.5. Les méthodes utilisées pour l'estimation régionale de la mobilité locale sont données dans la Section 5.6. La Section 5.7 regroupe les résultats des estimations réalisées. Une brève conclusion est donnée dans la Section 5.8.

**Mots clés** estimation sur petits domaines, estimation régionale, mobilité locale, transport

## 5.1 Contexte

Dans les pays développés, malgré le plafonnement observé dans les années 2000 (phénomène du *peak travel*[1]), la croissance des trafics a pour

[1]Le peak travel est une théorie selon laquelle le plafonnement de la mobilité des personnes (nombre de déplacements et distance parcourue par personne) correspondrait à un maximum.

conséquence directe l'augmentation des émissions de gaz à effet de serre. En France, depuis le "Grenelle Environnement" de 2007, les politiques souhaitent réduire les émissions de gaz à effet de serre des transports. En vue de prendre des décisions politiques plus durables, i.e. favorables à l'environnement, la collecte de données sur la mobilité est alors nécessaire afin d'étudier les comportements des personnes. En France, les Enquêtes Nationales sur les Transports (ENT) représentent un instrument essentiel de mesure de la mobilité. Les ENT sont réalisées auprès des ménages ordinaires résidant en France Métropolitaine. La dernière, l'Enquête Nationale sur les Transports et les Déplacements (ENTD) date de 2007-2008 ; c'est la cinquième de la lignée des ENT réalisées depuis les années 1960 (1966-1967, 1973-1974, 1981-1982 et 1993-1994). Elle en reprend d'ailleurs les définitions et les principes.

## 5.2 L'Enquête Nationale sur les Transports et les Déplacements (ENTD) 2007-2008

Cette Section reprend la description de l'ENTD 2007-2008 des notes méthodologiques de Roux & Armoogum (2008) et de Randrianasolo et al. (2010).

### 5.2.1 Présentation de l'ENTD

Cette présentation de l'ENTD reprend celle donnée dans les notes méthodologiques de Roux & Armoogum (2008) et de Randrianasolo et al. (2010).

*"L'objectif de cette enquête est la connaissance des déplacements des ménages résidant en France et de leur usage des moyens de transport tant collectifs qu'individuels. L'*ENTD *2007-2008 s'inscrit dans le prolongement des précédentes* ENT *afin d'assurer la mesure des évolutions structurelles.*

*Les principaux thèmes abordés sont les déplacements, de courtes et de longues distances des ménages et de leurs membres, selon les différents modes de transports. L'enquête met également l'accent sur la connaissance du parc de véhicules (voitures, motos, vélos, quads), de leur usage et l'accessibilité des individus aux transports collectifs (dont les abonnements et les réductions tarifaires).*

*L'*ENTD *est la seule sur la mobilité réalisée à cette échelle et qui décrit tous les déplacements, quels que soient le motif, la longueur, la durée, le mode de transport, la période de l'année ou le moment de la journée. Elle s'intéresse aussi aux possibilités d'accès aux transports collectifs et aux moyens de transports individuels dont disposent les ménages pour comprendre leurs comportements.*

*Comme les précédentes* ENT, *l'*ENTD *2007-2008 est organisée autour des trois grands thèmes suivants :*

- *Description des déplacements :*
  - *La mobilité régulière : les déplacements habituels domicile-travail, domicile-lieu d'étude ou de garde des enfants ;*
  - *La mobilité locale : les déplacements réalisés à l'occasion d'activités situées dans un rayon de 80 kilomètres autour du domicile ;*
  - *La mobilité à longues distances : les déplacements réalisés à plus de 80 kilomètres du lieu de sa résidence principale.*
- *Connaissance du parc de véhicules et de leurs usages :*
  - *Description des véhicules dont dispose le ménage et historique du parc du ménage (sur trois ans) ;*
  - *Description des usages d'un véhicule pendant une semaine ;*
  - *Détention de permis de conduire, pratique de la conduite et accidents de la circulation.*
- *Accessibilité aux transports collectifs (dont abonnements et réductions tarifaires)."*

La Figure 5.1 donne un aperçu de l'architecture de l'ENTD 2007-2008.

### 5.2.2 Le déroulement de l'ENTD

Cette description du déroulement de l'ENTD reprend celle donnée dans les notes méthodologiques de Roux & Armoogum (2008) et de Randrianasolo et al. (2010).

*"La phase de collecte des données de l'*ENTD *2007-2008 a été réalisée du 30 mars 2007 au 27 avril 2008 auprès des ménages ordinaires résidant en France métropolitaine. Elle s'est déroulée en six vagues successives (d'environ huit semaines chacune) sur tout le territoire métropolitain français de façon à éliminer les effets de saisonnalité (voyage à plus de 100 km plus nombreux en été, etc.) qui affectent la mobilité, tout au moins privée, des ménages. Le ménage reçoit l'enquêteur au cours de deux visites espacées de huit jours. Il y a donc deux interviews en face-à-face d'une durée totale d'environ deux heures. Les ménages sont interrogés dans leur résidence principale.*

*La première visite, orientée sur les facteurs explicatifs du comportement de la mobilité, autorise un* proxy, *i.e. un membre du ménage peut répondre à la place d'un autre membre absent lors de la visite de l'enquêteur.*

*A la fin de cette visite, un tirage au sort est effectué pour désigner, d'une*

*part, l'individu qui répondra aux questionnaires de la seconde visite et, d'autre part, un véhicule. Le ou les utilisateurs de ce dernier doivent noter tous les déplacements réalisés avec ce véhicule pendant sept jours consécutifs. Dans certains cas, l'individu tiré au sort, s'il est majeur et volontaire, reçoit un récepteur* Gps *à porter lors de tous ses déplacements au cours des sept prochains jours.*

*La seconde visite, orientée sur la description des comportements, n'autorise pas le proxy, sauf dans le cas d'un enfant de moins de 12 ans ou de personne inapte à répondre. Au cours de cette visite, le carnet véhicule ainsi que le* Gps *sont récupérés par l'enquêteur. Un module de questions supplémentaires est alors posé aux utilisateurs du* Gps *pour connaître les motifs et les moyens utilisés pour les déplacements et dans le cas contraire, les raisons de l'immobilité.*

*Lors de la visite initiale, l'enquête aborde, au niveau du ménage ou de l'ensemble de ses membres, les thèmes suivants :*

- *la description des caractéristiques socio-démographiques des individus qui composent le ménage ;*
- *les déplacements réguliers domicile-travail, domicile-lieu d'étude, domicile-lieu de garde des enfants ;*
- *le permis de conduire et pratique de la conduite, accidents de la circulation ;*
- *les abonnements et réductions dans les transports collectifs ;*
- *la description des véhicules dont dispose le ménage ;*
- *l'environnement du logement.*

*Lors de la seconde visite, l'individu tiré au sort (appelé individu* Kish*) répond aux questions portant sur la description :*

- *de ses déplacements de la veille (le jour de semaine à décrire est le jour le plus proche de la deuxième visite durant lequel le* Kish *est sorti de chez lui) et du dernier week-end (seule une journée de week-end est décrite, sauf dans deux régions à extension : Ile de France et Pays de Loire. L'enquêté* Kish *répond à des questions sur un jour où il a été mobile, le samedi ou le dimanche, avec une probabilité égale de tirage s'il n'est pas resté chez lui l'un de ces deux jours) ;*
- *de ses déplacements à longues distances pendant les trois derniers mois précédents la date de l'enquête. La période d'observation est réduite à deux mois s'il y a plus de dix voyages à décrire. Cette description des voyages fait appel à une mémoire des enquêtés plus lointaine qui peut être source d'oubli ou d'erreur ;*

- *des modes de transports employés tout au long de sa vie à l'aide d'une grille biographique (seulement en vagues 4 et 5, i.e. du 29 octobre 2007 au 2 mars 2008).*

*Entre les deux visites :*

- *un carnet véhicule a été tenu par le(s) conducteur(s) de l'un des véhicules du ménage pendant sept jours ;*
- *l'individu* Kish *a utilisé, uniquement s'il le souhaitait, un récepteur* GPS *au cours de ces déplacements pendant sept jours.*

*La taille de l'échantillon des répondants de l'*ENTD *2007-2008 est d'environ 20 200 ménages (y compris les sur-échantillons régionaux), la collecte a été étalée sur 12 mois (six vagues : 30 avril – 24 juin 2007 ; 25 juin – 2 septembre 2007 ; 3 septembre – 28 octobre 2007 ; 29 octobre – 23 décembre 2007 ; 2 janvier – 2 mars 2008 ; 3 mars – 27 avril 2008) afin de prendre en compte la saisonnalité qui marque la mobilité (surtout pour les déplacements à longue distance)."*

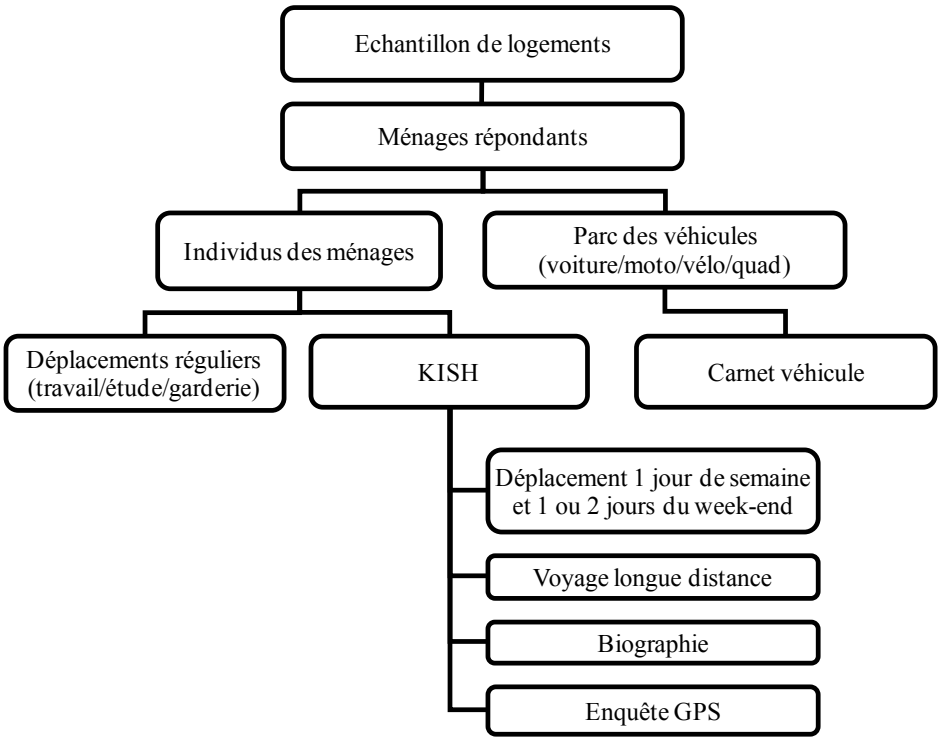


FIGURE 5.1 – *L'Enquête Nationale sur les Transports et les Déplacements (*ENTD*) 2007-2008. Figure tirée de l'Encadré 2 de Roux & Armoogum (2008) et du Schéma 1 de Randrianasolo et al. (2010).*

### 5.2.3 Le tirage de l'individu *Kish*

Cette description du tirage de l'individu *Kish* reprend celle donnée dans les notes méthodologiques de Roux & Armoogum (2008) et de Randrianasolo et al. (2010).

*"En fin de première visite, un individu du ménage (appelé individu* Kish*) est tiré au sort pour répondre aux questionnaires de la seconde visite parmi les personnes éligibles du ménage (individus de 6 ans et plus, présents au moment de la seconde visite et aptes à répondre). Ce tirage est à probabilités inégales afin d'interroger de préférence la personne qui fait le plus de voyages à longues distances."*

### 5.2.4 Le tirage du *carnet véhicule*

Cette description du tirage du *carnet véhicule* reprend celle donnée dans les notes méthodologiques de Roux & Armoogum (2008) et de Randrianasolo et al. (2010).

*"Un second tirage est effectué pour désigner le véhicule qui recevra un carnet (appelé* carnet véhicule*) à l'intérieur duquel tous les déplacements effectués avec ce véhicule seront consignés. Le tirage vise à privilégier d'abord les deux roues à moteur pour lesquels les informations sont rares, puis les vélos, puis les autres véhicules à moteur, sous la condition qu'ils aient été utilisés au cours du dernier mois. Les véhicules qui n'ont pas été utilisés au cours des quatre dernières semaines ont une probabilité plus faible ou nulle d'être tirés."*

## 5.3 Plan de sondage de l'ENTD 2007-2008

Cette Section reprend la description du plan de sondage de l'ENTD 2007-2008 des notes méthodologiques de Roux & Armoogum (2008, 2010) et de Randrianasolo et al. (2010).

### 5.3.1 La construction de l'échantillon national des ménages

Cette description de la construction de l'échantillon national des ménages reprend celle des notes méthodologiques de Roux & Armoogum (2008, 2010) et de Randrianasolo et al. (2010).

*"Le champ de l'enquête est celui des déplacements et du parc automobile des résidents de France métropolitaine. Les unités enquêtées sont les unités de vie telles qu'elles sont définies dans le tableau de composition du ménage (*TCM*) des enquêtes ménages de l'*INSEE*, i.e. l'ensemble des personnes ayant une résidence habituelle commune et partageant un même budget.*

*Le plan de sondage de l'*ENTD *2007-2008 est réalisé en deux phases afin de collecter le plus d'information possible sur les déplacements."*

### 5.3.2 Le plan de sondage de l'échantillon national

Cette description du plan de sondage de l'échantillon national reprend celle des notes méthodologiques de Roux & Armoogum (2008, 2010) et de Randrianasolo et al. (2010).

*"Au niveau national, l'échantillon de première phase de l'*ENTD *2007-2008 provient de l'échantillon-maître (EM) issu du recensement de la population de 1999, complété par la base de sondage des logements neufs achevés depuis le recensement de 1999. Les ménages multi-motorisés étant les plus mobiles, ces derniers ont été sur-représentés lors de la seconde phase du plan de sondage, de même que ceux habitant dans des communes rurales, qui sont rarement couvertes par les enquêtes locales (Enquêtes Ménage Déplacement coordonnées par le CERTU). L'échantillon de deuxième phase pour le niveau national regroupe 17500 logements, ce qui donne un taux de sondage moyen de 0,056%."*

#### L'échantillon-maître (EM) de 1999

La description de l'EM de 1999 donnée ci-après est largement basée sur l'article de Wilms (2000) intitulé *L'Échantillon-Maître 99 et Application au tirage des unités primaires par la macro "Cube"* et présenté aux Journées de Méthodologie Statistique de 2000.

Wilms (2000) définit l'échantillon-maître (EM) comme *"une réserve de logements destinée à alimenter la plupart des enquêtes-ménages nationales de l'*INSEE *entre deux recensements de la population"*. L'EM de 1999 contient donc des logements issus du recensement de 1999 et est disponible depuis 2001. Plus précisément, l'EM de 1999 contient environ 5% des logements recensés en 1999. L'EM de 1999 est le résultat d'un tirage stratifié à un ou deux degrés. La constitution de l'EM de 1999 se fait en deux étapes. La première étape consiste à partitionner la France Métropolitaine en cinq strates qui représentent des unités urbaines[2] ou des regroupements de communes rurales :

- strate 1 : regroupements de communes rurales,
- strate 2 : unités urbaines de moins de 20000 habitants,
- strate 3 : unités urbaines de 20000 à 100000 habitants,
- strate 4 : unités urbaines hors de Paris de plus de 100000 habitants,
- strate 5 : unités urbaines de Paris.

[2]Selon l'INSEE, *"une unité urbaine est une commune ou un ensemble de communes présentant une zone de bâti continu (pas de coupure de plus de 200 mètres entre deux constructions) qui compte au moins 2000 habitants."*

La deuxième étape consiste à tirer des districts ou des groupes de districts dans les strates 3, 4 et 5.

Dans les strates 1 et 2, les unités ne correspondent pas à des définitions administratives, certaines communes pouvant être très étendues ou très peu peuplées. Des unités dites *primaires* sont alors constituées au niveau de ces strates. Dans la strate 1, les unités primaires sont des regroupements de communes rurales contenant entre 1800 et 3600 logements principaux et devant appartenir à une même région. Dans la strate 2, les unités primaires sont des unités urbaines contenant au moins 1800 logements et où les unités urbaines regroupées doivent être situées géographiquement proches dans la même région. Au niveau de la strate 3, les unités primaires coïncident avec les unités urbaines. Les strates 1, 2 et 3 sont alors redécoupées en fonction des 22 régions pour donner 66 strates. Sous les contraintes du taux de sondage moyen pour une enquête nationale standard fixé à 1/2000 et du nombre moyen d'adresses par enquêteur fixé à 23, les nombres d'unités primaires tirées dans les strates 1, 2 et 3 sont alors fixés à 128, 75 et 93. Ainsi, au niveau de chacune des 66 strates, les unités primaires sont tirées à probabilités inégales proportionnelles aux nombres de résidences principales issus du recensement de 1999. Le tirage est équilibré sur quatre variables (revenu net imposable, les tranches d'âge 0-19 ans, 20-59 ans et 60 ans et plus) au niveau de huit super-régions[3] (Champagne-Ardenne/Nord-Pas-de-Calais/Lorraine/Alsace – Ile-de-France/Picardie/Haute-Normandie – Centre/Bourgogne/Franche-Comté – Auvergne/Languedoc-Roussillon/Limousin – Basse-Normandie/Bretagne – Rhône-Alpes/PACA/Corse – Aquitaine/Midi-Pyrénées – Pays de la Loire/Poitou-Charentes).

Au niveau de la strate 3, une nouvelle stratification de chacune des 93 unités urbaines tirées est réalisée par rapport au nombre des logements. 30 groupes de districts sont ensuite tirés à probabilités égales dans chacune des unités urbaines et dont la répartition par strate de ces 30 groupes se fait proportionnellement au nombre de résidences principales contenus dans les strates. Le tirage est équilibré sur les tranches d'âge 0-19 ans, 20-59 ans et 60 ans et plus.

Au niveau des strates 4 et 5, une nouvelle stratification des unités urbaines tirées est réalisée par rapport au nombre de districts. Le tirage des districts est un tirage stratifié à probabilités égales fixées à 3% au niveau des unités urbaines, et équilibré sur les tranches d'âge 0-19 ans, 20-59 ans et 60 ans et plus.

[3] Il est ici intéressant de remarquer que ces huit super-régions ne correspondent pas aux ZEAT (niveau 1 de la nomenclature Européenne) définies dans la note 1 de l'Introduction de ce manuscrit.

Finalement, l'EM de 1999 contient tous les logements des unités sélectionnées dans les strates 1 et 2 ainsi que les logements des districts et groupes de districts sélectionnés dans les strates 3, 4 et 5.

### 5.3.3 Le plan de sondage des extensions régionales

Cette description du plan de sondage des extensions régionales reprend celle des notes méthodologiques de Roux & Armoogum (2008, 2010) et de Randrianasolo et al. (2010).

*"En plus de l'échantillon national, cinq régions ont opté pour une extension régionale (Bretagne, Ile-de-France, Midi-Pyrénées, Languedoc-Roussillon et Pays de la Loire). Les coefficients de sur-représentation sont identiques à ceux du plan de sondage national. Ainsi, les extensions régionales ont ajouté 10308 logements à l'échantillon national avec la décomposition régionale suivante :*

- *686 logements en Bretagne ;*
- *7002 logements en Ile-de-France ;*
- *836 logements en Midi-Pyrénées ;*
- *1242 logements en Languedoc-Roussillon ;*
- *542 logements en Pays de la Loire."*

### 5.3.4 Le plan de sondage de l'extension locale du département de Loire-Atlantique

Cette description du plan de sondage de l'extension locale du département de Loire-Atlantique reprend celle des notes méthodologiques de Roux & Armoogum (2008, 2010) et de Randrianasolo et al. (2010).

*"Le département de Loire-Atlantique souhaitait obtenir des estimations de mobilité pour les sous-populations suivantes :*

- *Communauté urbaine de Nantes (CUN),*
- *Schéma de cohérence et d'organisation du territoire (SCOT) de Nantes-Saint Nazaire,*
- *Division territoriale d'aménagement (DTA),*
- *Le département de Loire-Atlantique (D44).*

*Ces territoires sont emboîtés, ce qui facilite la stratification de ces zones. Ainsi le pôle inter-régional d'ingénierie statistique ménages de l'INSEE a tiré 2066 logements supplémentaires par rapport aux tirages national et régional, dans*

*l'ensemble du département de Loire-Atlantique (en visant 1000 ménages répondants dans la CUN, en incluant tous les échantillons tirés (part de l'échantillon national tiré dans l'échantillon-maître et situé dans la CUN, part de l'échantillon EMEX [complémentaire de l'échantillon-maître national pour les extensions régionales] appartenant à la CUN, échantillon complémentaire local).*

*La base de sondage était constituée du recensement de 1999 et de la base de logements neufs (BLN) mise à jour en mai 2006, i.e. incluant les logements achevés à la fin du 1er trimestre 2006. L'échantillon a été réparti entre les deux au* prorata *du nombre de logements dans chaque base à l'intérieur de chaque zone.*

*Au total, 30165 fiches adresses ont été tirées de la base de sondage pour donner 20178 ménages et 18632 individus* Kish." Dans la suite de ce chapitre, les unités statistiques considérées sont les individus *Kish*. Le Tableau 5.1 donne les effectifs des individus *Kish* par région dans l'ENTD 2007-2008 ainsi que les nombres d'individus de 6 ans et plus appartenant à des ménages au niveau de chacune des régions fournis par les bases de données infracommunales[4] du recensement de la population de 2008.

## 5.4 Pondération de l'ENTD 2007-2008

La pondération de l'ENTD 2007-2008 a effectuée par Roux & Armoogum (2008, 2010); Armoogum & Roux (2012). Les marges utilisées pour la pondération des deux visites sont issues des données de l'Enquête Emploi effectuée par l'INSEE. Les structures des variables de calage sont identiques à celles du recensement rénové (pour plus de détails sur les structures des variables, voir Roux & Armoogum, 2010). Cette section reprend essentiellement la Section 4 de Armoogum & Roux (2012) qui elle-même reprend essentiellement la Section 3.3 de Roux & Armoogum (2008) et les Sections 3.1 et 3.2 de Roux & Armoogum (2010). Cette section énumère les variables de calage utilisées par ces derniers afin d'obtenir les pondérations finales au niveau ménages-individus et au niveau des individus *Kish*.

### 5.4.1 Pondération ménages-individus

#### Niveau national

Les variables de calage utilisées pour les deux visites sont :

- la catégorie socio-professionnelle de la personne de référence du ménage,

[4] Ces bases de données infracommunales sont disponibles sur le site Internet de l'INSEE dont le lien permanent est donné par http://www.insee.fr/fr/bases-de-donnees/default.asp?page=recensement/resultats/2008/donnees-detaillees-recensement-2008.htm.

TABLE 5.1 – *Effectifs de l'échantillon des individus* Kish *de l'*ENTD *2007-2008 et nombre d'individus de 6 ans et plus appartenant à des ménages en 2008, par région.*

| | Nombre d'individus *Kish* | Nombre d'individus de 6 ans et plus appartenant à des ménages |
|---|---|---|
| Ile-de-France | 5454 | 10456408 |
| Champagne-Ardenne | 320 | 1209086 |
| Picardie | 410 | 1711411 |
| Haute-Normandie | 228 | 1647437 |
| Centre | 449 | 2283546 |
| Basse-Normandie | 347 | 1325305 |
| Bourgogne | 387 | 1483458 |
| Nord | 608 | 3627613 |
| Lorraine | 442 | 2123618 |
| Alsace | 279 | 1664301 |
| Franche-Comté | 289 | 1051222 |
| Pays de la Loire | 2676 | 3146679 |
| Bretagne | 1139 | 2838958 |
| Poitou-Charente | 430 | 1591969 |
| Aquitaine | 708 | 2893526 |
| Midi-Pyrénées | 985 | 2578560 |
| Limousin | 206 | 674164 |
| Rhône-Alpes | 892 | 5505915 |
| Auvergne | 284 | 1215700 |
| Languedoc-Roussillon | 1294 | 2352795 |
| PACA | 758 | 4440467 |
| Corse | 47 | 279118 |

Source : ENTD 2007-2008, Bases de données infracommunales du Recensement de la population de 2008 de l'INSEE.

- l'âge croisé avec le sexe de la personne de référence du ménage,
- le type du ménage,
- le type d'immeuble,
- la nationalité de la personne de référence du ménage,
- la zone de résidence des ménages,
- le nombre d'individus (âge croisé avec le sexe),
- la motorisation du ménage,
- la vague de l'enquête (six vagues).

**Niveau régionale : Ile-de-France**

Les variables de calage utilisées sont :

- l'âge de la personne de référence du ménage,

- le sexe de la personne de référence du ménage,
- le type du ménage,
- le type d'immeuble,
- la catégorie socio-professionnelle de la personne de référence du ménage,
- la nationalité de la personne de référence,
- le nombre de ménages dans chaque département,
- la motorisation du ménage à Paris,
- la motorisation du ménage en Petite Couronne,
- la motorisation du ménage en Grande Couronne.

**Niveau régionale : Bretagne, Pays de la Loire, Midi-Pyrénées, Languedoc-Roussillon**

Les variables de calage utilisées pour chacune des régions sont :

- l'âge de la personne de référence du ménage,
- le sexe de la personne de référence du ménage,
- le type du ménage,
- le type d'immeuble,
- la catégorie socio-professionnelle de la personne de référence du ménage,
- la zone de résidence des ménages,
- la motorisation du ménage.

**Niveau locale des Pays de la Loire : Département de Loire-Atlantique, Communauté urbaine de Nantes, Division territoriale d'aménagement, Schéma de cohérence et d'organisation du territoire**

Les variables de calage utilisées pour chacune des localités sont :

- l'âge de la personne de référence,
- le sexe de la personne de référence.

### 5.4.2 Pondération individu *Kish*

La pondération initiale des individus *Kish* considérée par Roux & Armoogum (2008, 2010); Armoogum & Roux (2012) est celle des individus (pondération finale des ménages-individus en visite 2) divisée par la probabilité d'être tiré au sort pour répondre à ce volet de l'enquête.

**Niveau national**

Les variables de calage utilisées sont :

- la catégorie socio-professionnelle de l'individu *Kish*,
- l'âge croisé avec le sexe de l'individu *Kish*,
- la taille du ménage de l'individu *Kish*,
- la zone de résidence de l'individu *Kish*,
- la motorisation de l'individu *Kish*,
- la vague de l'enquête (six vagues),
- le jour de l'enquête (les jours ouvrés de la semaine).

**Niveau régionale : Ile-de-France**

Les variables de calage utilisées sont :

- la catégorie socio-professionnelle de l'individu *Kish*,
- l'âge croisé avec le sexe de l'individu *Kish*,
- la motorisation de l'individu *Kish*,
- le nombre d'individus de 6 ans et plus dans chaque département,
- la motorisation des individus de 6 ans et plus à Paris,
- la motorisation des individus de 6 ans et plus en Petite Couronne,
- la motorisation des individus de 6 ans et plus en Grande Couronne.

**Niveau régionale : Bretagne, Pays de la Loire, Midi-Pyrénées, Languedoc-Roussillon**

Les variables de calage utilisées pour chacune des régions sont :

- la catégorie socio-professionnelle de l'individu *Kish*,
- l'âge croisé avec le sexe de l'individu *Kish*, la zone de résidence de l'individu *Kish*,
- la motorisation de l'individu *Kish*.

**Niveau locale des Pays de la Loire : Département de Loire-Atlantique, Communauté urbaine de Nantes, Division territoriale d'aménagement, Schéma de cohérence et d'organisation du territoire**

Les variables de calage utilisées pour chacune des localités sont :

- l'âge de l'individu *Kish*,
- le sexe de l'individu *Kish*.

## 5.5 Objectif de l'étude

Le but de cette étude est de connaître la mobilité locale des individus de 6 ans et plus, un jour ouvré de la semaine i.e. du lundi au vendredi, en France Métropolitaine. Un déplacement est considéré comme local lorsque celui-ci est effectué dans un rayon de 80 km du domicile de l'individu *Kish* interrogé. Cette étude vise à estimer les indicateurs classiques de la mobilité locale au niveau de chaque région. Ces indicateurs classiques sont le nombre de déplacements, la distance parcourue ainsi que la durée des déplacements, par jour et par personne.

Un déplacement est un trajet ou un ensemble de trajets caractérisé par un motif à l'origine du déplacement et par un autre motif à l'arrivée du déplacement. Le trajet ou l'ensemble des trajets réalisé pour se rendre du domicile au travail est un exemple de déplacement. Dans le cas de l'ENTD 2007-2008, seules les distances des déplacements effectués en voiture ont été collectées. En effet, afin d'alléger le questionnaire, seuls les automobilistes ont été amenés à répondre à cette question. Les distances des déplacements effectués avec un autre mode de transport que la voiture ont donc été imputées. Certaines distances de déplacements effectuées en voiture ont été également imputées pour cause de non-réponse. Les imputations des distances des déplacements locaux ont été effectuées par Armoogum et al. (2012, p. 28). La méthode d'imputation a consisté à estimer les distances manquantes en considérant la durée du déplacement, la distance à vol d'oiseau entre les chefs-lieux des communes d'origine et de destination, tout en faisant attention à bien distinguer les communes voisines et les autres couples de communes. Pour chacun des modes de transports utilisés, des classes (croisements de variables) sont définies et un modèle est calculé au niveau de chaque classe en tenant compte des données de l'ENT 1993-1994. Malgré l'existence de ces imputations dans les données des individus *Kish*, celles-ci sont par la suite supposées certaines.

## 5.6 Méthodes d'estimation pour petits domaines

Une première étude sur les indicateurs régionaux de la mobilité a été réalisée par Le Guennec (2012) en comparant trois estimateurs pour petits domaines : l'estimateur redressé par calage régional, l'estimateur synthétique de type régression et un estimateur indirect basé sur une modélisation explicite sur les unités enquêtées (qui sont dans le cas de cette étude, les individus *Kish*), le meilleur estimateur empirique linéaire sans biais ou estimateur EBLUP. Les descriptions de ces méthodes sont données dans le Chapitre 1.2.

### 5.6.1 Méthodes utilisées

Ce chapitre propose des comparaisons de plusieurs estimateurs régionaux suivant deux grandes familles d'estimateurs pour petits domaines, essentiellement présentées dans la Section 1.2 du Chapitre 1.

#### Famille d'estimateurs pour petits domaines utilisant l'approche basée sur le plan de sondage

La première famille d'estimateurs utilise l'approche basée sur le plan de sondage et regroupe les estimateurs directs et les estimateurs indirects synthétiques et composites (voir Chapitre 1.2.1). Les estimateurs directs considérés sont l'estimateur de Horvitz & Thompson (1952), l'estimateur redressé par calage utilisant la pondération nationale de l'ENTD 2007-2008 et les estimateurs redressés par calage régional utilisant la méthode linéaire (avec une pseudo-distance de type *Khi-deux*), la méthode du raking ratio (avec une pseudo-distance de type *Entropie*) et la méthode logistique (avec une fonction de pseudo-distance de type logistique, voir Chapitre 2.2.1).

La méthode proposée dans le Chapitre 2, concernant la procédure d'optimisation du redressement d'un sous-échantillon d'une enquête, peut être appliquée à l'estimation régionale de la mobilité de l'individu *Kish*. La ventilation de l'échantillon par région peut en effet être considérée comme un ensemble de sous-échantillons disjoints de l'ENTD. Dans l'ENTD, le poids initial de l'individu *Kish* dépend directement de la taille du ménage auquel il appartient ainsi que de sa mobilité (les nombres de ses voyages de longues distances). De ce fait, les poids initiaux des individus *Kish* sont très disparates et peuvent être très grands. Un calage régional risque donc de faire disperser encore plus les poids de calage. Par ailleurs, dans le processus de redressement par calage régional, en relâchant les contraintes de calage, les résidus (définis dans le Chapitre 1.2.1) peuvent devenir très grands en valeur absolue. Tout repose donc sur un choix judicieux des variables auxiliaires utilisées. Des variables auxi-

liaires restreintes mentionnées plus tard dans la première partie de la Section 5.6.2 permettent d'obtenir des estimations par calage régional assez satisfaisantes. Les mêmes variables auxiliaires sont utilisées pour les estimations synthétiques et composites, afin de pouvoir comparer les résultats obtenus.

#### Famille d'estimateurs pour petits domaines utilisant l'approche basée sur une modélisation explicite

La seconde famille d'estimateurs pour petits domaines regroupe les estimateurs construits explicitement sur une modélisation (voir Chapitre 1.2.2). Dans la suite de ce chapitre, la modélisation se fera uniquement au niveau des unités d'observation, i.e. au niveau des individus *Kish*, car la France Métropolitaine ne dispose que de 22 régions et les unités d'observation sont disponibles au niveau de chaque région. Le modèle considéré sera alors toujours le modèle de régression linéaire à erreurs emboîtées. Par ailleurs, l'estimateur BLUP considéré dans la suite de ce chapitre sera celui proposé par Henderson (1975), de même donc pour l'estimateur pseudo-BLUP. Les estimateurs modifiés EBLUP et pseudo-EBLUP, proposés dans le Chapitre 4, intègrent également la seconde famille d'estimateurs. Enfin, bien que l'estimateur par scission des poids proposé dans le Chapitre 3.6 soit assisté par un modèle et non dépendant de celui-ci, l'estimateur est aussi regroupé dans la seconde famille d'estimateurs pour petits domaines. En effet, dans la construction de l'estimateur par scission, seule la constante de réglage (ou *tuning constant*) est empruntée au modèle de régression linéaire à erreurs emboîtées. Cependant, cette constante de réglage peut évidemment être choisie arbitrairement.

### 5.6.2 Variables auxiliaires

Plusieurs variables auxiliaires pour les 22 régions de la France Métropolitaine sont disponibles grâce aux données du recensement de 2008. Différents jeux de variables sont utilisés pour les deux grandes familles d'estimateurs. Les modalités des variables sont restreintes pour les estimations utilisant l'approche basée sur le plan. Pour les estimations utilisant l'approche basée explicitement sur un modèle, les modalités des variables sont plus détaillées.

#### Variables auxiliaires restreintes pour les estimateurs utilisant l'approche basée sur le plan

Les variables auxiliaires utilisées sont :

- le sexe de l'individu *Kish*,

- l'âge de l'individu *Kish* regroupé en 5 tranches (6-24 ans ; 25-34 ans ; 35-49 ans ; 50-64 ans ; 65 ans et plus),
- le type du ménage de l'individu *Kish* (personne seule ; famille monoparentale ; couple sans enfants ; couple avec enfants ; autres types de ménages),
- la catégorie socio-professionnelle de l'individu *Kish* (agriculteur, exploitant, artisan, commerçant, chef d'entreprise ; cadre, profession intellectuelle supérieure, profession intermédiaire ; employé ; ouvrier ; retraité ; autre sans activité professionnelle de 15 ans et plus ; individu de 6 à 15 ans),
- la zone de résidence de l'individu *Kish* (zone à dominante rurale, zone à dominante urbaine).

**Variables auxiliaires détaillées pour les estimateurs utilisant l'approche basée explicitement sur un modèle**

Les variables auxiliaires utilisées sont :

- le sexe de l'individu *Kish*,
- l'âge de l'individu *Kish* regroupé en 5 tranches (6-24 ans ; 25-34 ans ; 35-49 ans ; 50-64 ans ; 65 ans et plus),
- le type du ménage de l'individu *Kish* (personne seule ; famille monoparentale ; couple sans enfants ; couple avec enfants ; autres types de ménages),
- la catégorie socio-professionnelle de l'individu *Kish* (agriculteur, exploitant ; artisan, commerçant, chef d'entreprise ; cadre et profession intellectuelle supérieure ; profession intermédiaire ; employé ; ouvrier ; retraité ; autre sans activité professionnelle de 15 ans et plus ; individu de 6 à 15 ans),
- la zone de résidence de l'individu *Kish* (zone à dominante rurale, zone à dominante urbaine).

### 5.6.3 Plan de sondage national des individus *Kish*

Le plan de sondage national des individus *Kish* est à plusieurs degrés. Afin de simplifier, on peut considérer le plan de sondage comme un plan à trois degrés. Le premier degré est celui de l'EM de 1999 de l'Insee. Le second degré est un tirage de logements dans les unités primaires et dans les unités urbaines. Le second degré est en fait lui-même un plan de sondage à deux phases. La seconde phase consiste à retirer des logements dans la base de sondage, après stratification, de manière à surreprésenter

les ménages vivant en milieu rural et les ménages multi-motorisés. Le troisième degré est le tirage d'un individu par ménage, ou individu *Kish*, avec des probabilités inégales selon les nombres de voyages de longues distances.

Le plan de sondage au niveau national des individus *Kish* est donc très complexe. De plus, seules les probabilités d'inclusion d'ordre 1 des individus *Kish* sont disponibles dans les données sur la mobilité de l'ENTD 2007-2008. En effet, au niveau de chaque ménage, le tirage des individus *Kish* peut être vu comme un plan à probabilités inégales de taille fixe égale à 1, les probabilités d'inclusion d'ordre 2 sont donc nulles. Il n'existe donc pas d'estimateur sans biais. Tout ceci nous amène donc à approcher le plan de sondage des individus *Kish* par un plan de Poisson. Bien qu'il ne soit pas un plan de taille fixe, le plan de Poisson a tendance à surestimer la variance. Le choix du plan Poissonnien peut par conséquent être vu comme une stratégie conservatrice.

## 5.7 Résultats

Les estimations des indicateurs de mobilité obtenues suivant les différentes méthodes testées sont disponibles dans l'Annexe A.

### 5.7.1 Comparaison des méthodes

#### Estimations directes

Les résultats des estimations directes des indicateurs de mobilité locale régionale sont regroupés dans les Tableaux A.1, A.7 et A.13.

Au niveau des cinq régions ayant opté pour une extension (Ile-de-France, Pays de la Loire, Bretagne, Midi-Pyrénées et Languedoc-Roussillon), les estimations issues de la pondération nationale de l'ENTD sont très proches de celles issues des calages régionaux. De plus, ces estimations ont des précisions *quasi* similaires.

Au niveau des "petites" régions, les estimations issues des calages régionaux semblent meilleures que celles issues de la pondération nationale. Voire, les estimations issues de l'estimateur de Horvitz & Thompson (1952) semblent être meilleures que celles issues de la pondération nationale. Ceci pourrait être expliqué par le fait que la pondération nationale est issue d'un redressement par calage sur 240 modalités. La pondération a été effectuée afin d'améliorer la précision des estimations au niveau national et non nécessairement au niveau des régions (hormis les cinq régions ayant opté pour une extension). Le nombre assez important de

contraintes de calage (240 contraintes pour 18632 individus *Kish*), la forte dispersion des poids initiaux des individus *Kish* (principal défaut de la méthode *Kish* ; voir, par exemple, Deville, 1998), le calage effectué avec plusieurs variables catégorielles, provoquent nécessairement une explosion des poids de calage. Ce qui explique la diminution de la précision de certaines estimations régionales issues de la pondération nationale. L'utilisation de variables auxiliaires plus restreintes dans les calages régionaux a ainsi abouti à des estimations assez satisfaisantes. Cependant, il faudrait considérer les estimations obtenues par calages régionaux avec beaucoup de précaution, spécialement au niveau des "petites régions", car même si les contraintes de calage ont été relâchées, les estimations proviennent de calages qui sont restreints à chaque région et qui utilisent des poids initiaux très dispersés.

Etant donnée la faible taille du sous-échantillon représentant la Corse (47 individus *Kish*, voir Tableau 5.1), les calages régionaux utilisant la méthode du raking ratio et la méthode logistique ont été impossible à implémenter. En effet, les calages utilisant ces méthodes imposent une contrainte supplémentaire : les poids de calage doivent être positifs, et même bornés dans le cas de la méthode logistique.

**Estimations indirectes synthétiques et composites**

Les résultats des estimations indirectes synthétiques et composites des indicateurs de mobilité locale régionale sont respectivement regroupés dans les Tableaux A.2, A.8 et A.14, et dans les Tableaux A.3, A.9 et A.15.

Dans ce chapitre, les erreurs quadratiques moyennes des estimateurs synthétiques sont estimées selon l'approximation donnée dans le Chapitre 1.2.2. La formule d'approximation de l'erreur quadratique moyenne a la particularité d'inclure un estimateur direct de petits domaines. Trois approximations de l'erreur quadratique moyenne sont alors données (voir Tableaux A.2, A.8 et A.14). Ces trois approximations sont issues de trois choix de l'estimateur direct utilisé : l'estimateur direct issu de calages régionaux utilisant la méthode linéaire, la méthode du raking ratio et la méthode logistique. Une approximation utilisant l'estimateur direct construit avec les pondérations nationales a d'abord été tentée mais a conduit à des erreurs quadratiques moyennes négatives pour certaines régions et également à une moyenne négative des erreurs quadratiques moyennes.

Les trois approximations des erreurs quadratiques moyennes issues des trois estimateurs par calage régional sont également très instables (des valeurs négatives pour certaines régions). De ce fait, une moyenne des

erreurs quadratiques moyennes a été considérée. Les trois approximations des erreurs quadratiques moyennes de l'estimateur synthétique (présentées dans les Tableaux A.2, A.8 et A.14) ne sont donc plus spécifiques aux régions. Les erreurs sont plus faibles avec les estimateurs par calage régional utilisant la méthode du raking ratio et la méthode logistique car celles-ci sont la moyenne des erreurs quadratiques moyennes des 21 régions (la Corse étant exclue). De ce fait, les poids optimaux (ou constantes optimales de réglage) des estimations composites ne sont donc plus spécifiques aux régions mais uniques pour toutes les régions. Trois estimateurs composites optimaux sont alors également proposés dans les Tableaux A.3, A.9 et A.15.

**Estimations basées ou assistées par le modèle de régression linéaire à erreurs emboîtées**

Comme mentionné dans le Chapitre 3, la variance de l'estimateur par scission est très compliquée à calculer. En effet, l'estimateur par scission est un estimateur composite complexe doublement calé : calé à la fois sur les totaux des domaines et sur les totaux de la population globale. Sa variance peut cependant être estimée en ayant recours aux techniques de ré-échantillonnage. Il en est de même des variances des estimateurs Eblup et pseudo-Eblup modifiés décrits dans le Chapitre 4. Afin de pouvoir comparer les précisions de ces estimateurs avec celles des estimateurs Eblup et pseudo-Eblup classiques, toutes les estimations de variances considérées dans la suite de ce chapitre seront obtenues en utilisant les techniques de ré-échantillonnage.

Le plan de sondage national des individus *Kish* est supposé être un tirage Poissonnien comme mentionné dans la Section 5.6.3. Antal & Tillé (2011, Section 5) ont récemment proposé une procédure moins calculatoire de ré-échantillonnage pour un échantillon issu d'un tirage Poissonnien. L'idée principale de la méthode (voir Algorithme 5.1) consiste à sélectionner sans remise une partie des unités de l'échantillon suivant un tirage Poissonnien d'une part, et à sélectionner avec remise une partie des unités en utilisant des variables aléatoires de Poisson d'autre part.

Les résultats issus des estimations Eblup et pseudo-Eblup classiques et modifiées des indicateurs de mobilité locale par région, sont donnés dans les Tableaux A.4 et A.5, les Tableaux A.10 et A.11, et les Tableaux A.16 et A.17.

Les estimations Eblup classiques semblent être plus performantes que les estimations pseudo-Eblup classiques, notamment au niveau des grandes régions. Les performances sont assez équivalentes au niveau des petites

**Algorithme 5.1:** Procédure de ré-échantillonnage pour un échantillon issu d'un tirage Poissonnien, tirée de l'Algorithme 1 de Antal & Tillé (2011)

**pour** *chaque individu k de l'échantillon* **faire**
  Soit $\pi_k$ la probabilité d'inclusion de l'individu $k$ dans l'échantillon;
  Soit $u_k$ la valeur de la réalisation d'une variable aléatoire suivant la loi uniforme sur l'intervalle [0,1];
  Soit à initialiser $S^*_{kA} = 0$;
  **si** $u_k < \pi_k$ **alors**
    $S^*_{kA} = 1$
  Soit à initialiser $S^*_{kB} = 0$;
  **si** $S^*_{kA} = 1$ **alors**
    $S^*_{kB} = 0$
  **sinon**
    $S^*_{kB}$ prend la valeur de la réalisation d'une variable aléatoire suivant la loi de Poisson de paramètre $\lambda = 1$
  Soit $S^*_k = S^*_{kA} + S^*_{kB}$;
  Tirer l'individu $k$ $S^*_k$ fois dans l'échantillon;

régions, excepté au niveau de la Corse, où les estimations pseudo-Eblup classiques prennent le dessus sur les estimations Eblup classiques. Les constantes de réglage utilisées dans les estimations Eblup sont toujours plus élevées que celles utilisées dans les estimations pseudo-Eblup : les estimations Eblup donnent donc plus d'importance aux parties estimations directes que les estimations pseudo-Eblup, et inversement, elles donnent moins d'importance au modèle considéré que les estimations pseudo-Eblup.

Au niveau des régions ayant effectué une extension, la précision des estimations Eblup classiques et celles des estimations Eblup modifiées restent similaires. Au niveau des autres régions, les précisions des estimations Eblup modifiées sont légèrement meilleures, pour les variables *nombre de déplacements* et *durée totale de déplacements*. Inversement, la précision des estimations Eblup modifiées se dégrade pour la variable *distance totale*. Les estimations Eblup modifiées de type transfert semblent être légèrement plus performantes que celles des estimations Eblup modifiées de type compromis, dans le cas des variables *nombre de déplacements* et *durée totale de déplacements*. L'inverse se produit pour le cas de la variable *distance totale*.

Les comportements des estimations pseudo-Eblup modifiées par rapport

aux estimations pseudo-EBLUP classiques sont parallèlement similaires à ceux des estimations EBLUP.

Les résultats issus des estimations par scission des indicateurs de mobilité locale régionale, sont donnés dans les Tableaux A.6, A.12 et A.18.

Au niveau des régions ayant opté pour une extension, les précision des estimations par scission sont *quasi* similaires, peu importe le choix de la constante de réglage utilisée. Au niveau des autres régions, les estimations par scission utilisant des constantes de réglage provenant de l'estimateur pseudo-EBLUP semblent être légèrement plus performantes que celles utilisant des constantes provenant de l'estimateur EBLUP, pour les variables *nombre de déplacements* et *durée totale de déplacements*. C'est l'inverse qui se produit pour le cas de la variable *distance totale*. Similairement, au niveau des autres régions, dans le cas des variables *nombre de déplacements* et *durée totale de déplacements*, les estimations par scission utilisant une constante unique provenant de l'estimateur pseudo-EBLUP sont légèrement plus performantes que celles utilisant la constante unique provenant de l'estimateur EBLUP. Dans le cas de la variable *distance totale*, l'inverse se produit.

### 5.7.2 La mobilité locale au niveau des régions

Les estimations régionales des indicateurs de mobilité données dans l'Annexe A permettent de tirer quelques constatations sur les comportements de mobilité locale au niveau des régions.

Les résultats obtenus mettent en avant une grande spécificité de l'Ile-de-France en matière de mobilité locale : les personnes vivant en Ile-de-France figurent parmi celles qui se déplacent le moins, avec la distance totale parcourue la plus faible mais avec la durée totale des déplacements la plus élevée. Au niveau des autres régions, la durée totale des déplacements ainsi que la distance totale parcourue s'accroissent avec le nombre de déplacements.

## 5.8 Conclusion

Les estimations directes au niveau des régions ayant effectué des extensions sont assez précises en comparaison avec les estimations basées ou assistées par un modèle. Au niveau des autres régions, la précaution est de mise concernant les estimations directes, notamment à cause des poids des individus *Kish* qui peuvent être très grands. Afin de pallier à ces problèmes, Deville (1998) a proposé une alternative pour éviter des instabilités causées par les poids des individus *Kish* au niveau de leur

tirage. La précision des estimations synthétiques et composites s'avère également très instable pour l'estimation de la mobilité locale. Ceci peut s'expliquer par le choix de l'estimation directe disponible utilisée pour estimer leur précision.

Les estimations basées ou assistées par un modèle s'avèrent très précises et ont des performances *quasi* similaires d'une méthode à l'autre. Traditionnellement, les estimations basées sur un modèle fournissent uniquement des estimations, et non des pondérations susceptibles d'être réutilisées, ce qui fait l'avantage des nouvelles méthodes mises en œuvre dans ce chapitre (réécriture de l'Eblup et du pseudo Eblup, estimations Eblup et pseudo-Eblup modifiées, estimation par scission).

# Conclusion générale

Dans cette thèse, nous avons proposé des méthodes d'estimation sur petits domaines basées sur le plan de sondage. Ces méthodes mettent l'accent sur la construction de nouvelles pondérations pour chaque unité statistique. Tout au long de ce manuscrit, les méthodes proposées visent à fournir des estimations satisfaisant deux exigences : 1/ des estimations locales calées sur les totaux locaux des variables auxiliaires ; 2/ des estimations locales cohérentes avec l'estimation calée au niveau de l'ensemble de la population. Pour fournir des estimations calées, la variable d'intérêt doit être suffisamment corrélée avec les variables auxiliaires. Plus forte sera cette corrélation, meilleur sera le système de pondération obtenu.

La première méthode proposée dans ce manuscrit est une méthode directe qui fournit des estimations locales calées sur les totaux des variables auxiliaires dans chaque domaine. Il est à noter que la précision des estimations obtenues par calage local dépend clairement du choix des variables auxiliaires utilisées. C'est pourquoi, cette première méthode a cherché à optimiser, au moyen d'une procédure de sélection des variables auxiliaires pertinentes, la précision de l'estimation de la moyenne d'une variable d'intérêt. Bien évidemment, d'autres fonctions d'intérêt peuvent être considérées. Par ailleurs, cette méthode est susceptible de fournir des estimations des moyennes de plusieurs variables d'intérêt suffisamment précises en utilisant une pondération unique issue d'un calage sur des variables auxiliaires communes. Comme toute méthode directe, cette méthode proposée dans le Chapitre 2 fournit toutefois des estimations d'autant moins précises que la taille du domaine et faible.

La deuxième méthode proposée dans ce manuscrit est partie de l'idée d'emprunter de la force au niveau de tous les domaines : chaque pondération dépend à la fois d'une unité statistique et d'un domaine. Cette deuxième méthode proposée dans le Chapitre 3 a été implémentée en deux étapes. La première a consisté à caler les poids initiaux à la fois sur les totaux locaux des variables auxiliaires et sur les totaux des variables auxiliaires au niveau de l'ensemble de la population. L'estimation obte-

nue à cette première étape est de type synthétique, i.e. toutes les unités statistiques contribuent aux estimations au niveau de chaque domaine. La deuxième étape a alors consisté à combiner ce nouvel estimateur synthétique avec un estimateur direct puis de nouveau à doublement caler sur les totaux locaux des variables auxiliaires ainsi que sur les totaux au niveau de l'ensemble de la population. La moyenne de l'estimation synthétique avec l'estimation directe est pondérée par une constante de réglage qui dépend du domaine considéré. Cette constante de réglage peut être choisie arbitrairement. Toutefois, en ayant remarqué dans le Chapitre 1 que les estimations Eblup et pseudo-Eblup basées sur un modèle sont de type composite, i.e. une moyenne pondérée d'une estimation directe et d'une estimation synthétique, les mêmes constantes de réglages peuvent être utilisées dans la méthode proposée. Le Chapitre 3 montre que l'estimateur par scission ainsi obtenue se comporte de manière similaire aux estimateurs Eblup et pseudo-Eblup. La méthode proposée n'est cependant pas particulièrement robuste en cas de valeurs aberrantes. Notons que deux cas peuvent se présenter. Le premier cas est lorsque le couple (variable d'intérêt, variables auxiliaires) est aberrant et le second est lorsque seule la variable d'intérêt est aberrante. Le premier cas ne pose pas réellement problème car les poids de calage dépendent des variables auxiliaires et résout donc automatiquement le problème de valeurs aberrantes. Le second cas est un cas plus complexe qui pourrait être un sujet à approfondir.

La troisième méthode proposée dans ce manuscrit considère le comportement des estimateurs basés sur un modèle dans un environnement dont l'inférence est basée sur le plan de sondage. La méthode a consisté à réécrire les estimateurs Eblup et pseudo-Eblup en estimateurs homogènes linéaires, ce qui a alors permis d'obtenir des pondérations Eblup et pseudo-Eblup. Lorsque les pondérations Eblup sont appliquées aux variables auxiliaires, les propriétés de calage au niveau de chaque domaine sont satisfaites. Lorsque les pondérations pseudo-Eblup sont appliquées aux variables auxiliaires, les propriétés de calage au niveau de chaque domaine ainsi que la propriété de calage sur l'ensemble de la population sont satisfaites. La méthode montre que, pour plusieurs variables d'intérêt suffisamment corrélées entre elles, un système de pondération unique peut être obtenu. Les mêmes propriétés de calage sont conservées et la précision des nouvelles estimations Eblup et pseudo-Eblup dites modifiées, telles que nous les avons approchées par simulation avec application sur des données réelles, est assez proche de celles des estimateurs Eblup et pseudo-Eblup standards. Cependant, cette méthode proposée dans le Chapitre 4 fournit des pondérations pouvant prendre des valeurs négatives.

L'estimateur par scission et les estimateurs Eblup et pseudo-Eblup mo-

difiés proposés par les deux dernières méthodes sont très complexes : l'estimateur par scission est une moyenne pondérée doublement calée d'un estimateur direct calé et d'un estimateur synthétique doublement calé ; les estimateurs Eblup et pseudo-Eblup modifiés dépendent de paramètres qui sont la moyenne de quotients de variances. Les calculs des variances de ces estimateurs sont donc d'autant plus complexes à réaliser. Les estimateurs proposés étant obtenus à partir de méthodes pour petits domaines, il est clair que leurs variances devraient également être obtenues à partir de méthodes pour petits domaines. Toujours est-il que l'approximation de la variance au moyen des techniques de ré-échantillonnage semble être une bonne alternative.

La théorie générale de l'estimation par calage (dont l'estimation par la régression est un cas particulier) a été formalisée par Deville & Särndal (1992). L'estimateur par scission proposé dans le Chapitre 3 est un estimateur de type composite calé à la fois sur les totaux locaux des variables auxiliaires ainsi que sur les totaux de l'ensemble de la population. Pour cet estimateur, la procédure de calage est basée sur la méthode du raking ratio. Parallèlement, l'estimateur pseudo-Eblup est également un estimateur de type composite, "calé" à la fois sur les totaux locaux des variables auxiliaires ainsi que sur les totaux de l'ensemble de la population. Le calage de l'estimateur pseudo-Eblup peut être vu comme basé sur la méthode linéaire. Une question ouverte peut ainsi être posée : existerait-il une théorie d'hyper-calage dont l'estimateur par scission ainsi que l'estimateur pseudo-Eblup seraient des cas particuliers ? Au vu de l'analogie des comportements entre ces deux estimateurs, cette interrogation semble légitime.

Les trois méthodes proposées dans cette thèse ont été appliquées à l'estimation d'indicateurs de la mobilité locale à partir de l'Enquête Nationale sur les Transports et les Déplacements 2007-2008. Ces méthodes ont ainsi permis de mesurer la mobilité locale des personnes au niveau de chaque région de la France Métropolitaine. La précision des estimations obtenues reste satisfaisante même au niveau des régions dont les tailles sont faibles dans l'Enquête Nationale sur les Transports et les Déplacements 2007-2008.

# La mobilité locale par région en France Métropolitaine

A

Cette annexe regroupe les différents résultats de l'estimation régionale de la mobilité locale (c'est-à-dire à moins de 80 km du domicile) en France vue dans le Chapitre 5.

## A.1 Nombre de déplacements par personne un jour de semaine

TABLE A.1 – *Estimations directes du nombre de déplacements par personne un jour de semaine ainsi que les variances associées.*

| | | | Pondération | | Calage régional | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\pi$-estimateur | | nationale | | linéaire | | raking | | logit | |
| | est. | var. | est. | var. | est. | var. | est. | var. | est. | var. |
| Ile-de-France | 3.00 | 0.004 | 2.98 | 0.001 | 2.99 | 0.001 | 2.99 | 0.001 | 2.99 | 0.001 |
| Champagne-Ardenne | 2.94 | 0.056 | 2.94 | 0.075 | 3.02 | 0.033 | 3.01 | 0.034 | 3.01 | 0.034 |
| Picardie | 3.30 | 0.085 | 3.25 | 0.096 | 3.51 | 0.070 | 3.50 | 0.075 | 3.51 | 0.077 |
| Haute-Normandie | 3.24 | 0.070 | 3.11 | 0.092 | 3.45 | 0.071 | 3.42 | 0.070 | 3.42 | 0.070 |
| Centre | 2.87 | 0.032 | 2.96 | 0.047 | 2.89 | 0.024 | 2.89 | 0.023 | 2.87 | 0.023 |
| Basse-Normandie | 2.69 | 0.058 | 2.65 | 0.085 | 2.64 | 0.016 | 2.62 | 0.015 | 2.62 | 0.016 |
| Bourgogne | 3.30 | 0.082 | 3.25 | 0.110 | 3.26 | 0.032 | 3.26 | 0.033 | 3.26 | 0.033 |
| Nord | 3.37 | 0.059 | 3.27 | 0.061 | 3.31 | 0.023 | 3.31 | 0.022 | 3.30 | 0.022 |
| Lorraine | 3.04 | 0.055 | 2.96 | 0.053 | 3.02 | 0.022 | 3.04 | 0.023 | 3.04 | 0.023 |
| Alsace | 3.55 | 0.088 | 3.55 | 0.109 | 3.49 | 0.021 | 3.46 | 0.020 | 3.45 | 0.020 |
| Franche-Comté | 3.05 | 0.098 | 3.06 | 0.114 | 3.08 | 0.032 | 3.06 | 0.032 | 3.05 | 0.033 |
| Pays de la Loire | 3.49 | 0.019 | 3.52 | 0.006 | 3.51 | 0.006 | 3.51 | 0.006 | 3.51 | 0.006 |
| Bretagne | 3.51 | 0.033 | 3.51 | 0.008 | 3.51 | 0.008 | 3.51 | 0.008 | 3.51 | 0.008 |
| Poitou-Charente | 2.70 | 0.046 | 2.68 | 0.058 | 2.54 | 0.012 | 2.53 | 0.011 | 2.53 | 0.012 |
| Aquitaine | 2.77 | 0.030 | 2.86 | 0.041 | 2.92 | 0.013 | 2.93 | 0.013 | 2.94 | 0.013 |
| Midi-Pyrénées | 3.09 | 0.023 | 3.15 | 0.012 | 3.15 | 0.013 | 3.14 | 0.013 | 3.14 | 0.013 |
| Limousin | 3.12 | 0.124 | 3.27 | 0.190 | 3.19 | 0.045 | 3.17 | 0.045 | 3.17 | 0.045 |
| Rhône-Alpes | 3.25 | 0.020 | 3.27 | 0.027 | 3.28 | 0.010 | 3.28 | 0.010 | 3.28 | 0.010 |
| Auvergne | 3.14 | 0.090 | 3.26 | 0.121 | 3.14 | 0.032 | 3.15 | 0.033 | 3.15 | 0.033 |
| Languedoc-Roussillon | 3.23 | 0.018 | 3.19 | 0.007 | 3.19 | 0.009 | 3.19 | 0.009 | 3.19 | 0.009 |
| PACA | 3.32 | 0.035 | 3.30 | 0.047 | 3.32 | 0.016 | 3.31 | 0.016 | 3.31 | 0.016 |
| Corse | 2.73 | 0.316 | 3.07 | 0.600 | 2.71 | 0.053 | - | - | - | - |

Source : ENTD 2007-2008.

TABLE A.2 – *Estimations indirectes synthétiques du nombre de déplacements par personne un jour de semaine ainsi que trois estimations de l'erreur quadratique moyenne non spécifiques aux régions. Les trois estimations de l'erreur quadratique moyenne sont issues de trois choix de l'estimateur direct utilisé dans le calcul de la précision : l'estimateur direct issu d'un calage régional utilisant la méthode linéaire, la méthode du raking ratio et la méthode logistique. Les estimations des erreurs quadratiques moyennes présentées dans ce tableau sont obtenues par la formule de Gonzalez & Wakesberg (1973) qui donne la moyenne des erreurs quadratiques moyennes sur l'ensemble des régions. Contrairement aux estimations des erreurs quadratiques par région qui sont très instables et peuvent prendre des valeurs négatives pour certaines régions, ces estimations de l'erreur quadratique moyenne ont l'avantage d'être stables mais ont l'inconvénient de ne pas être spécifiques à chaque région.*

| | Estimation | Erreur quadratique moyenne | | |
|---|---|---|---|---|
| | | (1) | (2) | (3) |
| Ile-de-France | 3.27 | 0.043 | 0.040 | 0.041 |
| Champagne-Ardenne | 3.13 | // | // | // |
| Picardie | 3.18 | // | // | // |
| Haute-Normandie | 3.18 | // | // | // |
| Centre | 3.12 | // | // | // |
| Basse-Normandie | 3.10 | // | // | // |
| Bourgogne | 3.08 | // | // | // |
| Nord | 3.20 | // | // | // |
| Lorraine | 3.16 | // | // | // |
| Alsace | 3.21 | // | // | // |
| Franche-Comté | 3.13 | // | // | // |
| Pays de la Loire | 3.14 | // | // | // |
| Bretagne | 3.12 | // | // | // |
| Poitou-Charente | 3.07 | // | // | // |
| Aquitaine | 3.10 | // | // | // |
| Midi-Pyrénées | 3.11 | // | // | // |
| Limousin | 3.03 | // | // | // |
| Rhône-Alpes | 3.19 | // | // | // |
| Auvergne | 3.08 | // | // | // |
| Languedoc-Roussillon | 3.09 | // | // | // |
| PACA | 3.13 | // | // | // |
| Corse | 3.08 | 0.043 | *0.040* | *0.041* |

Source : ENTD 2007-2008.

TABLE A.3 – *Estimations indirectes composites du nombre de déplacements par personne un jour de semaine ainsi que trois estimations du poids optimal non spécifiques aux régions. Les trois estimations composites sont issues de trois choix de l'estimateur direct utilisé dans le calcul de la précision : l'estimateur direct issu d'un calage régional utilisant la méthode linéaire, la méthode du raking ratio ainsi que la méthode logistique. Les constantes de réglage utilisées minimisent les erreurs quadratiques moyennes des estimateurs. Comme les erreurs quadratiques moyennes des estimations synthétiques ne dépendent pas des régions, les constantes de réglage utilisées dans ce tableau ne dépendent pas non plus des régions.*

| | $\hat{\phi}^{*1}$ | est. (1) | $\hat{\phi}^{*2}$ | est. (2) | $\hat{\phi}^{*3}$ | est. (3) |
|---|---|---|---|---|---|---|
| Ile-de-France | 0.629 | 3.09 | 0.622 | 3.10 | 0.623 | 3.10 |
| Champagne-Ardenne | 〃 | 3.06 | 〃 | 3.05 | 〃 | 3.05 |
| Picardie | 〃 | 3.39 | 〃 | 3.38 | 〃 | 3.38 |
| Haute-Normandie | 〃 | 3.35 | 〃 | 3.33 | 〃 | 3.33 |
| Centre | 〃 | 2.97 | 〃 | 2.97 | 〃 | 2.96 |
| Basse-Normandie | 〃 | 2.81 | 〃 | 2.80 | 〃 | 2.80 |
| Bourgogne | 〃 | 3.19 | 〃 | 3.19 | 〃 | 3.19 |
| Nord | 〃 | 3.27 | 〃 | 3.27 | 〃 | 3.27 |
| Lorraine | 〃 | 3.07 | 〃 | 3.08 | 〃 | 3.09 |
| Alsace | 〃 | 3.38 | 〃 | 3.36 | 〃 | 3.36 |
| Franche-Comté | 〃 | 3.10 | 〃 | 3.09 | 〃 | 3.08 |
| Pays de la Loire | 〃 | 3.37 | 〃 | 3.37 | 〃 | 3.37 |
| Bretagne | 〃 | 3.36 | 〃 | 3.36 | 〃 | 3.36 |
| Poitou-Charente | 〃 | 2.73 | 〃 | 2.73 | 〃 | 2.73 |
| Aquitaine | 〃 | 2.99 | 〃 | 3.00 | 〃 | 3.00 |
| Midi-Pyrénées | 〃 | 3.13 | 〃 | 3.13 | 〃 | 3.13 |
| Limousin | 〃 | 3.13 | 〃 | 3.12 | 〃 | 3.12 |
| Rhône-Alpes | 〃 | 3.24 | 〃 | 3.25 | 〃 | 3.25 |
| Auvergne | 〃 | 3.12 | 〃 | 3.12 | 〃 | 3.12 |
| Languedoc-Roussillon | 〃 | 3.15 | 〃 | 3.15 | 〃 | 3.15 |
| PACA | 〃 | 3.25 | 〃 | 3.25 | 〃 | 3.25 |
| Corse | 0.629 | 2.85 | *0.622* | *3.08* | *0.623* | *3.08* |

Source : ENTD 2007-2008.

TABLE A.4 – *Estimations* EBLUP *standard et modifiées du nombre de déplacements par personne un jour de semaine. L'estimation* EBLUP *de type compromis utilise une constante de réglage (tuning constant) obtenue à partir des variables nombre de déplacements, distance totale parcourue et durée totale des déplacements. L'estimation* EBLUP *de type transfert utilise une constante de réglage obtenue à partir des variables distance totale parcourue et durée totale des déplacements.*

| | EBLUP | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Standard | | | Compromis | | | Transfert | | |
| | $\hat{\gamma}_i$ | est. | var. | $\hat{\Gamma}_i$ | est. | var. | $\hat{\Gamma}_i$ | est. | var. |
| Ile-de-France | 0.98 | 3.06 | 0.001 | 0.96 | 3.06 | 0.001 | 0.96 | 3.06 | 0.001 |
| Champagne-Ardenne | 0.73 | 3.12 | 0.011 | 0.65 | 3.12 | 0.010 | 0.61 | 3.12 | 0.009 |
| Picardie | 0.78 | 3.17 | 0.013 | 0.70 | 3.17 | 0.011 | 0.66 | 3.17 | 0.011 |
| Haute-Normandie | 0.66 | 3.08 | 0.012 | 0.57 | 3.10 | 0.010 | 0.53 | 3.10 | 0.009 |
| Centre | 0.79 | 2.86 | 0.009 | 0.71 | 2.88 | 0.008 | 0.68 | 2.90 | 0.007 |
| Basse-Normandie | 0.75 | 2.79 | 0.009 | 0.66 | 2.83 | 0.007 | 0.62 | 2.84 | 0.007 |
| Bourgogne | 0.77 | 3.17 | 0.010 | 0.69 | 3.16 | 0.008 | 0.65 | 3.15 | 0.008 |
| Nord | 0.84 | 3.22 | 0.007 | 0.77 | 3.22 | 0.006 | 0.73 | 3.22 | 0.006 |
| Lorraine | 0.79 | 3.02 | 0.010 | 0.71 | 3.03 | 0.008 | 0.67 | 3.04 | 0.008 |
| Alsace | 0.70 | 3.45 | 0.014 | 0.62 | 3.42 | 0.012 | 0.58 | 3.41 | 0.010 |
| Franche-Comté | 0.71 | 3.03 | 0.011 | 0.63 | 3.04 | 0.009 | 0.58 | 3.05 | 0.009 |
| Pays de la Loire | 0.96 | 3.47 | 0.003 | 0.93 | 3.46 | 0.002 | 0.92 | 3.46 | 0.002 |
| Bretagne | 0.91 | 3.42 | 0.004 | 0.86 | 3.41 | 0.004 | 0.83 | 3.40 | 0.004 |
| Poitou-Charente | 0.78 | 2.73 | 0.006 | 0.71 | 2.76 | 0.005 | 0.67 | 2.78 | 0.005 |
| Aquitaine | 0.86 | 2.97 | 0.007 | 0.79 | 2.98 | 0.006 | 0.76 | 2.98 | 0.006 |
| Midi-Pyrénées | 0.89 | 3.06 | 0.004 | 0.84 | 3.06 | 0.004 | 0.81 | 3.06 | 0.004 |
| Limousin | 0.64 | 3.09 | 0.014 | 0.55 | 3.08 | 0.011 | 0.51 | 3.08 | 0.010 |
| Rhône-Alpes | 0.88 | 3.20 | 0.005 | 0.83 | 3.20 | 0.005 | 0.80 | 3.20 | 0.004 |
| Auvergne | 0.71 | 3.10 | 0.011 | 0.62 | 3.09 | 0.009 | 0.58 | 3.09 | 0.009 |
| Languedoc-Roussillon | 0.92 | 3.25 | 0.003 | 0.87 | 3.24 | 0.003 | 0.85 | 3.23 | 0.003 |
| PACA | 0.87 | 3.19 | 0.006 | 0.80 | 3.18 | 0.006 | 0.77 | 3.18 | 0.006 |
| Corse | 0.28 | 2.87 | 0.008 | 0.24 | 2.90 | 0.006 | 0.21 | 2.92 | 0.005 |

Source : ENTD 2007-2008.

TABLE A.5 – *Estimations pseudo-*EBLUP *standard et modifiées du nombre de déplacements par personne un jour de semaine. L'estimation pseudo-*EBLUP *de type compromis utilise une constante de réglage (tuning constant) obtenue à partir des variables nombre de déplacements, distance totale parcourue et durée totale des déplacements. L'estimation pseudo-*EBLUP *de type transfert utilise une constante de réglage obtenue à partir des variables distance totale parcourue et durée totale des déplacements.*

| | Pseudo-EBLUP | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Standard | | | Compromis | | | Transfert | | |
| | $\hat{\gamma}_{iw}$ | est. | var. | $\hat{\Gamma}_{iw}$ | est. | var. | $\hat{\Gamma}_{iw}$ | est. | var. |
| Ile-de-France | 0.96 | 3.01 | 0.001 | 0.94 | 3.02 | 0.001 | 0.93 | 3.02 | 0.001 |
| Champagne-Ardenne | 0.61 | 3.02 | 0.013 | 0.52 | 3.03 | 0.011 | 0.48 | 3.04 | 0.010 |
| Picardie | 0.66 | 3.35 | 0.020 | 0.58 | 3.33 | 0.017 | 0.53 | 3.32 | 0.016 |
| Haute-Normandie | 0.52 | 3.24 | 0.024 | 0.44 | 3.23 | 0.019 | 0.40 | 3.23 | 0.017 |
| Centre | 0.67 | 2.92 | 0.012 | 0.58 | 2.94 | 0.010 | 0.54 | 2.96 | 0.009 |
| Basse-Normandie | 0.56 | 2.83 | 0.009 | 0.48 | 2.87 | 0.007 | 0.44 | 2.89 | 0.006 |
| Bourgogne | 0.60 | 3.17 | 0.015 | 0.52 | 3.16 | 0.013 | 0.47 | 3.16 | 0.011 |
| Nord | 0.76 | 3.31 | 0.017 | 0.68 | 3.30 | 0.015 | 0.64 | 3.30 | 0.014 |
| Lorraine | 0.69 | 3.08 | 0.013 | 0.60 | 3.09 | 0.011 | 0.56 | 3.10 | 0.010 |
| Alsace | 0.59 | 3.43 | 0.012 | 0.51 | 3.40 | 0.009 | 0.46 | 3.39 | 0.008 |
| Franche-Comté | 0.58 | 3.16 | 0.014 | 0.50 | 3.16 | 0.011 | 0.46 | 3.16 | 0.010 |
| Pays de la Loire | 0.89 | 3.46 | 0.006 | 0.84 | 3.44 | 0.005 | 0.81 | 3.44 | 0.005 |
| Bretagne | 0.84 | 3.45 | 0.007 | 0.77 | 3.43 | 0.006 | 0.73 | 3.41 | 0.006 |
| Poitou-Charente | 0.67 | 2.77 | 0.008 | 0.59 | 2.81 | 0.007 | 0.54 | 2.82 | 0.006 |
| Aquitaine | 0.77 | 2.90 | 0.007 | 0.69 | 2.93 | 0.006 | 0.65 | 2.94 | 0.006 |
| Midi-Pyrénées | 0.82 | 3.11 | 0.009 | 0.75 | 3.11 | 0.008 | 0.71 | 3.11 | 0.008 |
| Limousin | 0.50 | 3.11 | 0.015 | 0.42 | 3.10 | 0.012 | 0.38 | 3.10 | 0.011 |
| Rhône-Alpes | 0.80 | 3.24 | 0.006 | 0.73 | 3.24 | 0.005 | 0.69 | 3.24 | 0.005 |
| Auvergne | 0.55 | 3.09 | 0.015 | 0.47 | 3.09 | 0.012 | 0.43 | 3.09 | 0.011 |
| Languedoc-Roussillon | 0.85 | 3.22 | 0.007 | 0.79 | 3.22 | 0.007 | 0.76 | 3.21 | 0.006 |
| PACA | 0.78 | 3.30 | 0.012 | 0.70 | 3.29 | 0.010 | 0.66 | 3.28 | 0.010 |
| Corse | 0.17 | 3.02 | 0.007 | 0.14 | 3.04 | 0.005 | 0.12 | 3.04 | 0.005 |

Source : ENTD 2007-2008.

TABLE A.6 – *Estimations par scission du nombre de déplacements par personne un jour de semaine. Les constantes de réglages proviennent respectivement de l'estimateur* EBLUP *et de l'estimateur pseudo-*EBLUP*. La constante de réglage est dite spécifique lorsque celle-ci dépend uniquement du nombre de déplacement. Elle est dite unique lorsque celle-ci dépend des variables nombre de déplacements, distance totale parcourue et durée totale des déplacements.*

| | Constantes de réglage | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | tirées de l'EBLUP | | | | tirées du pseudo-EBLUP | | | |
| | spécifique | | unique | | spécifique | | unique | |
| | est. | var. | est. | var. | est. | var. | est. | var. |
| Ile-de-France | 3.03 | 0.002 | 3.04 | 0.002 | 3.04 | 0.002 | 3.05 | 0.002 |
| Champagne-Ardenne | 3.04 | 0.015 | 3.05 | 0.012 | 3.04 | 0.013 | 3.05 | 0.011 |
| Picardie | 3.31 | 0.023 | 3.30 | 0.019 | 3.30 | 0.021 | 3.29 | 0.017 |
| Haute-Normandie | 3.27 | 0.022 | 3.26 | 0.017 | 3.26 | 0.019 | 3.25 | 0.014 |
| Centre | 3.03 | 0.013 | 3.03 | 0.011 | 3.03 | 0.012 | 3.03 | 0.010 |
| Basse-Normandie | 2.80 | 0.010 | 2.86 | 0.008 | 2.82 | 0.009 | 2.89 | 0.007 |
| Bourgogne | 3.24 | 0.013 | 3.21 | 0.010 | 3.22 | 0.011 | 3.19 | 0.009 |
| Nord | 3.32 | 0.014 | 3.31 | 0.012 | 3.31 | 0.013 | 3.30 | 0.011 |
| Lorraine | 3.05 | 0.013 | 3.06 | 0.011 | 3.06 | 0.012 | 3.07 | 0.010 |
| Alsace | 3.40 | 0.012 | 3.37 | 0.010 | 3.38 | 0.010 | 3.35 | 0.008 |
| Franche-Comté | 3.07 | 0.015 | 3.08 | 0.012 | 3.08 | 0.013 | 3.09 | 0.010 |
| Pays de la Loire | 3.45 | 0.006 | 3.43 | 0.005 | 3.44 | 0.006 | 3.42 | 0.005 |
| Bretagne | 3.46 | 0.005 | 3.44 | 0.005 | 3.45 | 0.005 | 3.42 | 0.005 |
| Poitou-Charente | 2.74 | 0.009 | 2.79 | 0.008 | 2.77 | 0.009 | 2.82 | 0.007 |
| Aquitaine | 2.93 | 0.009 | 2.95 | 0.008 | 2.94 | 0.008 | 2.96 | 0.007 |
| Midi-Pyrénées | 3.09 | 0.009 | 3.10 | 0.008 | 3.09 | 0.009 | 3.10 | 0.008 |
| Limousin | 3.15 | 0.020 | 3.13 | 0.015 | 3.13 | 0.017 | 3.12 | 0.013 |
| Rhône-Alpes | 3.24 | 0.005 | 3.23 | 0.005 | 3.23 | 0.005 | 3.23 | 0.004 |
| Auvergne | 3.09 | 0.015 | 3.10 | 0.012 | 3.10 | 0.014 | 3.10 | 0.010 |
| Languedoc-Roussillon | 3.20 | 0.007 | 3.19 | 0.006 | 3.19 | 0.006 | 3.18 | 0.006 |
| PACA | 3.25 | 0.009 | 3.24 | 0.008 | 3.25 | 0.009 | 3.23 | 0.007 |
| Corse | 2.97 | 0.009 | 2.98 | 0.005 | 2.97 | 0.007 | 2.99 | 0.005 |

Source : ENTD 2007-2008.

## A.2 Distance totale parcourue par personne un jour de semaine

TABLE A.7 – *Estimations directes de la distance parcourue par personne un jour de semaine ainsi que les variances associées.*

| | | | Pondération | | Calage régional | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\pi$-estimateur | | nationale | | linéaire | | raking | | logit | |
| | est. | var. | est. | var. | est. | var. | est. | var. | est. | var. |
| Ile-de-France | 22.6 | 0.3 | 22.6 | 0.2 | 23.0 | 0.2 | 23.0 | 0.2 | 23.0 | 0.2 |
| Champagne-Ardenne | 23.3 | 4.3 | 23.9 | 6.0 | 23.5 | 3.5 | 23.6 | 3.6 | 23.6 | 3.6 |
| Picardie | 29.5 | 7.1 | 31.6 | 12.2 | 33.8 | 6.1 | 33.4 | 5.8 | 33.3 | 5.8 |
| Haute-Normandie | 26.6 | 9.2 | 28.4 | 22.7 | 27.9 | 7.4 | 27.5 | 7.7 | 27.5 | 7.7 |
| Centre | 23.2 | 3.4 | 24.0 | 5.4 | 26.2 | 5.1 | 26.1 | 5.1 | 26.2 | 5.2 |
| Basse-Normandie | 19.1 | 3.0 | 18.9 | 4.8 | 20.6 | 2.3 | 20.4 | 2.2 | 20.4 | 2.2 |
| Bourgogne | 21.5 | 5.6 | 21.4 | 6.4 | 22.9 | 3.5 | 23.0 | 3.5 | 23.0 | 3.5 |
| Nord | 24.8 | 3.9 | 25.1 | 5.5 | 24.8 | 2.4 | 24.6 | 2.3 | 24.6 | 2.3 |
| Lorraine | 25.6 | 5.5 | 26.9 | 7.4 | 27.9 | 4.8 | 28.2 | 4.9 | 28.3 | 5.0 |
| Alsace | 25.0 | 7.7 | 25.6 | 10.7 | 26.2 | 4.2 | 26.1 | 4.1 | 26.1 | 4.2 |
| Franche-Comté | 26.1 | 10.4 | 27.3 | 16.0 | 28.4 | 5.9 | 28.3 | 5.9 | 28.3 | 5.9 |
| Pays de la Loire | 27.6 | 1.6 | 27.8 | 0.8 | 28.0 | 0.9 | 28.0 | 0.9 | 28.0 | 0.9 |
| Bretagne | 28.1 | 2.8 | 28.8 | 1.4 | 29.1 | 1.4 | 29.1 | 1.4 | 29.0 | 1.4 |
| Poitou-Charente | 29.3 | 12.5 | 30.7 | 26.8 | 27.6 | 7.2 | 27.5 | 7.2 | 27.6 | 7.2 |
| Aquitaine | 23.2 | 2.8 | 24.3 | 4.2 | 24.1 | 2.3 | 24.3 | 2.3 | 24.4 | 2.3 |
| Midi-Pyrénées | 25.3 | 2.1 | 25.7 | 1.7 | 25.5 | 1.7 | 25.5 | 1.6 | 25.5 | 1.6 |
| Limousin | 28.7 | 14.7 | 30.6 | 22.1 | 29.7 | 9.9 | 29.7 | 9.7 | 29.8 | 9.7 |
| Rhône-Alpes | 24.7 | 2.0 | 25.3 | 3.1 | 26.3 | 3.0 | 26.4 | 3.2 | 26.4 | 3.2 |
| Auvergne | 25.5 | 10.1 | 24.6 | 10.5 | 24.7 | 3.9 | 24.9 | 3.9 | 25.0 | 3.9 |
| Languedoc-Roussillon | 24.0 | 1.9 | 23.9 | 1.5 | 24.1 | 1.5 | 24.2 | 1.5 | 24.2 | 1.5 |
| PACA | 25.1 | 3.0 | 25.9 | 5.1 | 26.2 | 2.9 | 26.2 | 3.0 | 26.2 | 3.0 |
| Corse | 18.5 | 37.8 | 21.3 | 86.2 | 16.6 | 17.6 | - | - | - | - |

Source : ENTD 2007-2008.

TABLE A.8 – *Estimations indirectes synthétiques de la distance parcourue par personne un jour de semaine ainsi que trois estimations de l'erreur quadratique moyenne non spécifiques aux régions. Les trois estimations de l'erreur quadratique moyenne sont issues de trois choix de l'estimateur direct utilisé dans le calcul de la précision : l'estimateur direct issu d'un calage régional utilisant la méthode linéaire, la méthode du raking ratio et la méthode logistique. Les estimations des erreurs quadratiques moyennes présentées dans ce tableau sont obtenues par la formule de Gonzalez & Wakesberg (1973) qui donne la moyenne des erreurs quadratiques moyennes sur l'ensemble des régions. Contrairement aux estimations des erreurs quadratiques par région qui sont très instables et peuvent prendre des valeurs négatives pour certaines régions, ces estimations de l'erreur quadratique moyenne ont l'avantage d'être stables mais ont l'inconvénient de ne pas être spécifiques à chaque région.*

| | Estimation | Erreur quadratique moyenne | | |
|---|---|---|---|---|
| | | (1) | (2) | (3) |
| Ile-de-France | 25.8 | 7.5 | 4.5 | 4.5 |
| Champagne-Ardenne | 25.7 | 〃 | 〃 | 〃 |
| Picardie | 25.7 | 〃 | 〃 | 〃 |
| Haute-Normandie | 24.9 | 〃 | 〃 | 〃 |
| Centre | 25.8 | 〃 | 〃 | 〃 |
| Basse-Normandie | 26.0 | 〃 | 〃 | 〃 |
| Bourgogne | 25.7 | 〃 | 〃 | 〃 |
| Nord | 24.2 | 〃 | 〃 | 〃 |
| Lorraine | 25.1 | 〃 | 〃 | 〃 |
| Alsace | 25.2 | 〃 | 〃 | 〃 |
| Franche-Comté | 25.7 | 〃 | 〃 | 〃 |
| Pays de la Loire | 26.3 | 〃 | 〃 | 〃 |
| Bretagne | 25.8 | 〃 | 〃 | 〃 |
| Poitou-Charente | 25.9 | 〃 | 〃 | 〃 |
| Aquitaine | 25.7 | 〃 | 〃 | 〃 |
| Midi-Pyrénées | 26.2 | 〃 | 〃 | 〃 |
| Limousin | 25.7 | 〃 | 〃 | 〃 |
| Rhône-Alpes | 25.6 | 〃 | 〃 | 〃 |
| Auvergne | 25.9 | 〃 | 〃 | 〃 |
| Languedoc-Roussillon | 25.2 | 〃 | 〃 | 〃 |
| PACA | 24.2 | 〃 | 〃 | 〃 |
| Corse | 25.5 | 7.5 | *4.5* | *4.5* |

Source : ENTD 2007-2008.

TABLE A.9 – *Estimations indirectes composites de la distance parcourue par personne un jour de semaine ainsi que trois estimations du poids optimal non spécifiques aux régions. Les trois estimations composites sont issues de trois choix de l'estimateur direct utilisé dans le calcul de la précision : l'estimateur direct issu d'un calage régional utilisant la méthode linéaire, la méthode du raking ratio et la méthode logistique. Les constantes de réglage utilisées minimisent les erreurs quadratiques moyennes des estimateurs. Comme les erreurs quadratiques moyennes des estimations synthétiques ne dépendent pas des régions, les constantes de réglage utilisées dans ce tableau ne dépendent pas non plus des régions.*

| | $\hat{\phi}^{*1}$ | est. (1) | $\hat{\phi}^{*2}$ | est. (2) | $\hat{\phi}^{*3}$ | est. (3) |
|---|---|---|---|---|---|---|
| Ile-de-France | 0.626 | 24.1 | 0.541 | 24.3 | 0.538 | 24.3 |
| Champagne-Ardenne | 〃 | 24.3 | 〃 | 24.5 | 〃 | 24.6 |
| Picardie | 〃 | 30.8 | 〃 | 29.8 | 〃 | 29.8 |
| Haute-Normandie | 〃 | 26.8 | 〃 | 26.3 | 〃 | 26.3 |
| Centre | 〃 | 26.0 | 〃 | 26.0 | 〃 | 26.0 |
| Basse-Normandie | 〃 | 22.6 | 〃 | 23.0 | 〃 | 23.0 |
| Bourgogne | 〃 | 23.9 | 〃 | 24.2 | 〃 | 24.3 |
| Nord | 〃 | 24.6 | 〃 | 24.4 | 〃 | 24.4 |
| Lorraine | 〃 | 26.9 | 〃 | 26.8 | 〃 | 26.8 |
| Alsace | 〃 | 25.8 | 〃 | 25.7 | 〃 | 25.7 |
| Franche-Comté | 〃 | 27.4 | 〃 | 27.1 | 〃 | 27.1 |
| Pays de la Loire | 〃 | 27.3 | 〃 | 27.2 | 〃 | 27.2 |
| Bretagne | 〃 | 27.8 | 〃 | 27.5 | 〃 | 27.5 |
| Poitou-Charente | 〃 | 27.0 | 〃 | 26.8 | 〃 | 26.8 |
| Aquitaine | 〃 | 24.7 | 〃 | 25.0 | 〃 | 25.0 |
| Midi-Pyrénées | 〃 | 25.7 | 〃 | 25.8 | 〃 | 25.8 |
| Limousin | 〃 | 28.2 | 〃 | 27.9 | 〃 | 27.9 |
| Rhône-Alpes | 〃 | 26.0 | 〃 | 26.1 | 〃 | 26.0 |
| Auvergne | 〃 | 25.1 | 〃 | 25.4 | 〃 | 25.4 |
| Languedoc-Roussillon | 〃 | 24.5 | 〃 | 24.6 | 〃 | 24.7 |
| PACA | 〃 | 25.4 | 〃 | 25.3 | 〃 | 25.3 |
| Corse | 0.626 | 19.9 | *0.541* | *25.5* | *0.538* | *25.5* |

Source : ENTD 2007-2008.

TABLE A.10 – *Estimations* EBLUP *standard et modifiées de la distance parcourue par personne un jour de semaine. L'estimation* EBLUP *de type compromis utilise une constante de réglage (tuning constant) obtenue à partir des variables nombre de déplacements, distance totale parcourue et durée totale des déplacements. L'estimation* EBLUP *de type transfert utilise une constante de réglage obtenue à partir des variables nombre de déplacements et durée totale des déplacements.*

| | EBLUP | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Standard | | | Compromis | | | Transfert | | |
| | $\hat{\gamma}_i$ | est. | var. | $\hat{\Gamma}_i$ | est. | var. | $\hat{\Gamma}_i$ | est. | var. |
| Ile-de-France | 0.93 | 25.1 | 0.1 | 0.96 | 25.0 | 0.1 | 0.98 | 25.0 | 0.1 |
| Champagne-Ardenne | 0.45 | 28.1 | 1.1 | 0.65 | 28.3 | 1.6 | 0.74 | 28.4 | 1.8 |
| Picardie | 0.51 | 30.2 | 2.3 | 0.70 | 31.1 | 2.4 | 0.79 | 31.5 | 2.5 |
| Haute-Normandie | 0.37 | 27.3 | 0.9 | 0.57 | 27.3 | 1.5 | 0.68 | 27.3 | 1.8 |
| Centre | 0.54 | 28.1 | 1.1 | 0.71 | 28.2 | 1.5 | 0.80 | 28.2 | 1.8 |
| Basse-Normandie | 0.47 | 26.5 | 1.0 | 0.66 | 25.9 | 1.3 | 0.76 | 25.7 | 1.5 |
| Bourgogne | 0.50 | 25.9 | 1.0 | 0.69 | 25.3 | 1.2 | 0.78 | 24.9 | 1.4 |
| Nord | 0.61 | 26.8 | 0.7 | 0.77 | 26.8 | 0.9 | 0.85 | 26.8 | 1.0 |
| Lorraine | 0.53 | 28.6 | 1.8 | 0.71 | 29.0 | 2.3 | 0.80 | 29.1 | 2.6 |
| Alsace | 0.42 | 27.4 | 1.0 | 0.62 | 27.2 | 1.5 | 0.72 | 27.1 | 1.8 |
| Franche-Comté | 0.43 | 27.7 | 1.0 | 0.63 | 27.7 | 1.5 | 0.72 | 27.6 | 1.8 |
| Pays de la Loire | 0.87 | 28.8 | 0.3 | 0.93 | 28.8 | 0.3 | 0.96 | 28.8 | 0.4 |
| Bretagne | 0.75 | 29.3 | 0.6 | 0.86 | 29.6 | 0.7 | 0.91 | 29.7 | 0.7 |
| Poitou-Charente | 0.53 | 28.0 | 1.2 | 0.71 | 28.1 | 1.6 | 0.80 | 28.1 | 1.8 |
| Aquitaine | 0.65 | 27.7 | 1.3 | 0.79 | 27.7 | 1.5 | 0.87 | 27.7 | 1.6 |
| Midi-Pyrénées | 0.72 | 27.8 | 0.7 | 0.84 | 27.8 | 0.8 | 0.90 | 27.7 | 0.9 |
| Limousin | 0.35 | 29.4 | 2.5 | 0.55 | 30.5 | 3.2 | 0.65 | 31.1 | 3.7 |
| Rhône-Alpes | 0.70 | 28.1 | 0.9 | 0.83 | 28.2 | 1.1 | 0.89 | 28.2 | 1.1 |
| Auvergne | 0.42 | 27.9 | 1.3 | 0.62 | 28.0 | 2.0 | 0.72 | 28.1 | 2.3 |
| Languedoc-Roussillon | 0.77 | 26.0 | 0.7 | 0.87 | 25.9 | 0.7 | 0.92 | 25.8 | 0.8 |
| PACA | 0.66 | 26.5 | 0.9 | 0.80 | 26.5 | 1.1 | 0.87 | 26.5 | 1.2 |
| Corse | 0.11 | 26.4 | 0.7 | 0.24 | 25.4 | 1.2 | 0.30 | 24.9 | 1.7 |

Source : ENTD 2007-2008.

TABLE A.11 – *Estimations pseudo-*EBLUP *standard et modifiées de la distance parcourue par personne un jour de semaine. L'estimation pseudo-*EBLUP *de type compromis utilise une constante de réglage (tuning constant) obtenue à partir des variables nombre de déplacements, distance totale parcourue et durée totale des déplacements. L'estimation pseudo-*EBLUP *de type transfert utilise une constante de réglage obtenue à partir des variables nombre de déplacements et durée totale des déplacements.*

| | Pseudo-EBLUP | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Standard | | | Compromis | | | Transfert | | |
| | $\hat{\gamma}_{iw}$ | est. | var. | $\hat{\Gamma}_{iw}$ | est. | var. | $\hat{\Gamma}_{iw}$ | est. | var. |
| Ile-de-France | 0.89 | 23.4 | 0.2 | 0.94 | 23.2 | 0.2 | 0.97 | 23.1 | 0.2 |
| Champagne-Ardenne | 0.32 | 25.3 | 1.0 | 0.52 | 24.8 | 1.5 | 0.63 | 24.6 | 1.8 |
| Picardie | 0.37 | 28.0 | 2.1 | 0.58 | 29.0 | 2.4 | 0.68 | 29.5 | 2.6 |
| Haute-Normandie | 0.25 | 26.0 | 2.0 | 0.44 | 26.4 | 3.4 | 0.54 | 26.7 | 4.4 |
| Centre | 0.38 | 25.8 | 1.2 | 0.58 | 25.5 | 1.8 | 0.68 | 25.4 | 2.2 |
| Basse-Normandie | 0.28 | 24.6 | 0.8 | 0.48 | 23.4 | 0.9 | 0.57 | 22.8 | 1.1 |
| Bourgogne | 0.31 | 25.1 | 1.2 | 0.52 | 24.5 | 1.6 | 0.62 | 24.2 | 1.9 |
| Nord | 0.49 | 24.8 | 0.9 | 0.68 | 24.8 | 1.2 | 0.77 | 24.8 | 1.4 |
| Lorraine | 0.40 | 26.4 | 1.5 | 0.60 | 26.8 | 2.0 | 0.70 | 27.0 | 2.3 |
| Alsace | 0.30 | 25.8 | 0.9 | 0.51 | 25.8 | 1.4 | 0.61 | 25.8 | 1.8 |
| Franche-Comté | 0.30 | 26.8 | 1.0 | 0.50 | 27.2 | 1.6 | 0.60 | 27.5 | 2.0 |
| Pays de la Loire | 0.71 | 27.8 | 0.6 | 0.84 | 28.0 | 0.6 | 0.90 | 28.1 | 0.7 |
| Bretagne | 0.61 | 27.7 | 1.0 | 0.77 | 28.1 | 1.2 | 0.85 | 28.3 | 1.3 |
| Poitou-Charente | 0.38 | 27.3 | 2.9 | 0.59 | 27.8 | 4.0 | 0.69 | 28.1 | 4.7 |
| Aquitaine | 0.50 | 25.2 | 1.1 | 0.69 | 24.9 | 1.4 | 0.78 | 24.7 | 1.5 |
| Midi-Pyrénées | 0.58 | 26.0 | 0.9 | 0.75 | 25.9 | 1.1 | 0.83 | 25.8 | 1.2 |
| Limousin | 0.23 | 26.7 | 2.6 | 0.42 | 27.2 | 4.0 | 0.52 | 27.5 | 5.0 |
| Rhône-Alpes | 0.55 | 25.7 | 1.1 | 0.73 | 25.5 | 1.3 | 0.81 | 25.5 | 1.5 |
| Auvergne | 0.27 | 26.1 | 1.4 | 0.47 | 26.1 | 2.2 | 0.57 | 26.1 | 2.7 |
| Languedoc-Roussillon | 0.64 | 25.2 | 0.9 | 0.79 | 25.0 | 1.0 | 0.86 | 25.0 | 1.1 |
| PACA | 0.51 | 25.6 | 1.0 | 0.70 | 25.9 | 1.3 | 0.79 | 26.0 | 1.5 |
| Corse | 0.06 | 25.7 | 0.6 | 0.14 | 25.4 | 1.1 | 0.18 | 25.3 | 1.6 |

Source : ENTD 2007-2008.

TABLE A.12 – *Estimations par scission de la distance parcourue par personne un jour de semaine. Les constantes de réglages proviennent respectivement de l'estimateur* EBLUP *et de l'estimateur pseudo-*EBLUP*. La constante de réglage est dite spécifique lorsque celle-ci dépend uniquement de la distance parcourue par personne un jour de semaine. Elle est dite unique lorsque celle-ci dépend des variables nombre de déplacements, distance totale parcourue et durée totale des déplacements.*

| | Constantes de réglage | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | tirées de l'EBLUP | | | | tirées du pseudo-EBLUP | | | |
| | spécifique | | unique | | spécifique | | unique | |
| | est. | var. | est. | var. | est. | var. | est. | var. |
| Ile-de-France | 23.3 | 0.3 | 23.5 | 0.3 | 23.1 | 0.3 | 23.2 | 0.3 |
| Champagne-Ardenne | 24.8 | 1.4 | 25.1 | 1.0 | 24.4 | 1.7 | 24.6 | 1.3 |
| Picardie | 29.4 | 2.6 | 28.5 | 2.1 | 30.2 | 2.7 | 29.6 | 2.2 |
| Haute-Normandie | 25.6 | 1.0 | 25.5 | 0.6 | 25.7 | 1.4 | 25.6 | 1.0 |
| Centre | 26.2 | 1.9 | 26.3 | 1.3 | 26.0 | 2.3 | 26.2 | 1.8 |
| Basse-Normandie | 23.8 | 1.1 | 24.6 | 0.8 | 23.1 | 1.3 | 23.8 | 0.9 |
| Bourgogne | 24.3 | 1.2 | 24.8 | 0.8 | 23.8 | 1.4 | 24.2 | 1.0 |
| Nord | 24.8 | 1.0 | 24.6 | 0.8 | 24.9 | 1.1 | 24.8 | 1.0 |
| Lorraine | 26.7 | 2.0 | 26.4 | 1.5 | 27.0 | 2.4 | 26.8 | 1.9 |
| Alsace | 25.8 | 1.4 | 25.7 | 1.0 | 25.8 | 1.9 | 25.8 | 1.5 |
| Franche-Comté | 27.5 | 1.8 | 27.1 | 1.2 | 27.9 | 2.4 | 27.6 | 1.8 |
| Pays de la Loire | 27.5 | 0.6 | 27.4 | 0.5 | 27.7 | 0.7 | 27.6 | 0.5 |
| Bretagne | 28.4 | 1.1 | 28.0 | 0.9 | 28.8 | 1.1 | 28.5 | 1.0 |
| Poitou-Charente | 27.1 | 2.7 | 26.9 | 2.1 | 27.2 | 3.2 | 27.2 | 2.7 |
| Aquitaine | 25.3 | 1.4 | 25.4 | 1.1 | 25.2 | 1.7 | 25.2 | 1.4 |
| Midi-Pyrénées | 25.7 | 0.9 | 25.8 | 0.7 | 25.5 | 1.1 | 25.6 | 0.9 |
| Limousin | 27.8 | 4.0 | 26.9 | 2.8 | 28.9 | 5.0 | 28.1 | 3.7 |
| Rhône-Alpes | 25.9 | 1.3 | 26.0 | 1.1 | 25.8 | 1.5 | 25.9 | 1.3 |
| Auvergne | 25.9 | 1.6 | 26.0 | 1.1 | 25.6 | 2.1 | 25.8 | 1.6 |
| Languedoc-Roussillon | 24.6 | 0.9 | 24.8 | 0.7 | 24.6 | 1.0 | 24.6 | 0.8 |
| PACA | 25.6 | 1.1 | 25.4 | 0.9 | 25.8 | 1.3 | 25.6 | 1.1 |
| Corse | 24.9 | 0.9 | 25.5 | 0.7 | 23.9 | 1.1 | 24.6 | 0.8 |

Source : ENTD 2007-2008.

## A.3 Durée totale des déplacements par personne un jour de semaine

TABLE A.13 – *Estimations directes de la durée totale des déplacements par personne un jour de semaine et les variances associées.*

| | | | Pondération | | Calage régional | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\pi$-estimateur | | nationale | | linéaire | | raking | | logit | |
| | est. | var. | est. | var. | est. | var. | est. | var. | est. | var. |
| Ile-de-France | 72 | 2 | 71 | 1 | 73 | 1 | 73 | 1 | 73 | 1 |
| Champagne-Ardenne | 46 | 15 | 47 | 20 | 46 | 9 | 46 | 9 | 47 | 9 |
| Picardie | 55 | 22 | 57 | 31 | 59 | 16 | 59 | 16 | 59 | 16 |
| Haute-Normandie | 53 | 19 | 53 | 35 | 55 | 17 | 54 | 17 | 54 | 17 |
| Centre | 45 | 9 | 47 | 13 | 48 | 8 | 47 | 8 | 47 | 8 |
| Basse-Normandie | 40 | 14 | 40 | 22 | 42 | 7 | 41 | 7 | 41 | 7 |
| Bourgogne | 49 | 23 | 49 | 32 | 49 | 9 | 48 | 8 | 48 | 8 |
| Nord | 52 | 12 | 52 | 15 | 52 | 5 | 52 | 5 | 52 | 5 |
| Lorraine | 50 | 15 | 51 | 17 | 53 | 10 | 53 | 10 | 53 | 10 |
| Alsace | 60 | 26 | 60 | 32 | 60 | 9 | 60 | 9 | 60 | 9 |
| Franche-Comté | 52 | 37 | 54 | 47 | 56 | 15 | 55 | 15 | 55 | 15 |
| Pays de la Loire | 54 | 4 | 54 | 2 | 55 | 2 | 55 | 2 | 55 | 2 |
| Bretagne | 54 | 9 | 55 | 3 | 55 | 3 | 55 | 3 | 55 | 3 |
| Poitou-Charente | 47 | 21 | 48 | 38 | 45 | 10 | 45 | 10 | 45 | 10 |
| Aquitaine | 47 | 9 | 49 | 13 | 49 | 7 | 50 | 7 | 50 | 7 |
| Midi-Pyrénées | 52 | 6 | 52 | 4 | 52 | 4 | 52 | 4 | 52 | 4 |
| Limousin | 51 | 34 | 52 | 49 | 51 | 17 | 51 | 17 | 51 | 17 |
| Rhône-Alpes | 54 | 6 | 54 | 9 | 57 | 6 | 57 | 6 | 57 | 6 |
| Auvergne | 48 | 22 | 47 | 24 | 47 | 11 | 47 | 11 | 47 | 11 |
| Languedoc-Roussillon | 53 | 5 | 53 | 3 | 53 | 3 | 53 | 3 | 53 | 3 |
| PACA | 65 | 18 | 63 | 21 | 66 | 18 | 66 | 17 | 66 | 17 |
| Corse | 45 | 86 | 51 | 187 | 69 | 63 | - | - | - | - |

Source : ENTD 2007-2008.

TABLE A.14 – *Estimations indirectes synthétiques de la durée totale des déplacements par personne un jour de semaine ainsi que trois estimations de l'erreur quadratique moyenne non spécifiques aux régions. Les trois estimations de l'erreur quadratique moyenne sont issues de trois choix de l'estimateur direct utilisé dans le calcul de la précision : l'estimateur direct issu d'un calage régional utilisant la méthode linéaire, la méthode du raking ratio ainsi que la méthode logistique. Les estimations des erreurs quadratiques moyennes présentées dans ce tableau sont obtenues par la formule de Gonzalez & Wakesberg (1973) qui donne la moyenne des erreurs quadratiques moyennes sur l'ensemble des régions. Contrairement aux estimations des erreurs quadratiques par région qui sont très instables et peuvent prendre des valeurs négatives pour certaines régions, ces estimations de l'erreur quadratique moyenne ont l'avantage d'être stables mais ont l'inconvénient de ne pas être spécifiques à chaque région.*

| | Estimation | Erreur quadratique moyenne | | |
|---|---|---|---|---|
| | | (1) | (2) | (3) |
| Ile-de-France | 61 | 38 | 35 | 35 |
| Champagne-Ardenne | 56 | 〃 | 〃 | 〃 |
| Picardie | 57 | 〃 | 〃 | 〃 |
| Haute-Normandie | 57 | 〃 | 〃 | 〃 |
| Centre | 56 | 〃 | 〃 | 〃 |
| Basse-Normandie | 55 | 〃 | 〃 | 〃 |
| Bourgogne | 55 | 〃 | 〃 | 〃 |
| Nord | 58 | 〃 | 〃 | 〃 |
| Lorraine | 57 | 〃 | 〃 | 〃 |
| Alsace | 58 | 〃 | 〃 | 〃 |
| Franche-Comté | 56 | 〃 | 〃 | 〃 |
| Pays de la Loire | 56 | 〃 | 〃 | 〃 |
| Bretagne | 56 | 〃 | 〃 | 〃 |
| Poitou-Charente | 55 | 〃 | 〃 | 〃 |
| Aquitaine | 56 | 〃 | 〃 | 〃 |
| Midi-Pyrénées | 56 | 〃 | 〃 | 〃 |
| Limousin | 54 | 〃 | 〃 | 〃 |
| Rhône-Alpes | 58 | 〃 | 〃 | 〃 |
| Auvergne | 55 | 〃 | 〃 | 〃 |
| Languedoc-Roussillon | 55 | 〃 | 〃 | 〃 |
| PACA | 57 | 〃 | 〃 | 〃 |
| Corse | 54 | 38 | *35* | *35* |

Source : ENTD 2007-2008.

TABLE A.15 – *Estimations indirectes composites de la durée totale des déplacements par personne un jour de semaine ainsi que trois estimations du poids optimal non spécifiques aux régions. Les trois estimations composites sont issues de trois choix de l'estimateur direct utilisé dans le calcul de la précision : l'estimateur direct issu d'un calage régional utilisant la méthode linéaire, la méthode du raking ratio ainsi que la méthode logistique. Les constantes de réglage utilisées minimisent les erreurs quadratiques moyennes des estimateurs. Comme les erreurs quadratiques moyennes des estimations synthétiques ne dépendent pas des régions, les constantes de réglage utilisées dans ce tableau ne dépendent pas non plus des régions.*

| | $\hat{\phi}^{*1}$ | est. (1) | $\hat{\phi}^{*2}$ | est. (2) | $\hat{\phi}^{*3}$ | est. (3) |
|---|---|---|---|---|---|---|
| Ile-de-France | 0.773 | 70 | 0.800 | 70 | 0.798 | 70 |
| Champagne-Ardenne | 〃 | 49 | 〃 | 48 | 〃 | 48 |
| Picardie | 〃 | 59 | 〃 | 59 | 〃 | 59 |
| Haute-Normandie | 〃 | 55 | 〃 | 54 | 〃 | 54 |
| Centre | 〃 | 50 | 〃 | 49 | 〃 | 49 |
| Basse-Normandie | 〃 | 45 | 〃 | 44 | 〃 | 44 |
| Bourgogne | 〃 | 50 | 〃 | 50 | 〃 | 50 |
| Nord | 〃 | 53 | 〃 | 53 | 〃 | 53 |
| Lorraine | 〃 | 54 | 〃 | 54 | 〃 | 54 |
| Alsace | 〃 | 60 | 〃 | 59 | 〃 | 59 |
| Franche-Comté | 〃 | 56 | 〃 | 55 | 〃 | 55 |
| Pays de la Loire | 〃 | 55 | 〃 | 55 | 〃 | 55 |
| Bretagne | 〃 | 55 | 〃 | 55 | 〃 | 55 |
| Poitou-Charente | 〃 | 47 | 〃 | 47 | 〃 | 47 |
| Aquitaine | 〃 | 51 | 〃 | 51 | 〃 | 51 |
| Midi-Pyrénées | 〃 | 53 | 〃 | 53 | 〃 | 53 |
| Limousin | 〃 | 52 | 〃 | 52 | 〃 | 52 |
| Rhône-Alpes | 〃 | 57 | 〃 | 57 | 〃 | 57 |
| Auvergne | 〃 | 49 | 〃 | 49 | 〃 | 49 |
| Languedoc-Roussillon | 〃 | 53 | 〃 | 54 | 〃 | 54 |
| PACA | 〃 | 64 | 〃 | 64 | 〃 | 64 |
| Corse | 0.773 | 66 | *0.800* | *54* | *0.798* | *54* |

Source : ENTD 2007-2008.

TABLE A.16 – *Estimations* EBLUP *standard et modifiées de la durée totale des déplacements par personne un jour de semaine. L'estimation* EBLUP *de type compromis utilise une constante de réglage (tuning constant) obtenue à partir des variables nombre de déplacements, distance totale parcourue et durée totale des déplacements. L'estimation* EBLUP *de type transfert utilise une constante de réglage obtenue à partir des variables nombre de déplacements et distance totale parcourue.*

| | EBLUP | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Standard | | | Compromis | | | Transfert | | |
| | $\hat{\gamma}_i$ | est. | var. | $\hat{\Gamma}_i$ | est. | var. | $\hat{\Gamma}_i$ | est. | var. |
| Ile-de-France | 0.98 | 73 | 1 | 0.96 | 72 | 1 | 0.96 | 72 | 1 |
| Champagne-Ardenne | 0.76 | 51 | 3 | 0.65 | 51 | 3 | 0.59 | 52 | 3 |
| Picardie | 0.80 | 56 | 5 | 0.70 | 56 | 4 | 0.65 | 56 | 4 |
| Haute-Normandie | 0.69 | 51 | 3 | 0.57 | 52 | 3 | 0.51 | 52 | 2 |
| Centre | 0.82 | 49 | 4 | 0.71 | 50 | 3 | 0.66 | 50 | 3 |
| Basse-Normandie | 0.77 | 47 | 4 | 0.66 | 48 | 4 | 0.61 | 49 | 3 |
| Bourgogne | 0.79 | 48 | 3 | 0.69 | 49 | 3 | 0.63 | 49 | 3 |
| Nord | 0.86 | 53 | 2 | 0.77 | 53 | 2 | 0.72 | 53 | 2 |
| Lorraine | 0.81 | 53 | 5 | 0.71 | 53 | 4 | 0.66 | 53 | 4 |
| Alsace | 0.73 | 59 | 4 | 0.62 | 58 | 3 | 0.56 | 58 | 3 |
| Franche-Comté | 0.74 | 50 | 5 | 0.63 | 51 | 4 | 0.57 | 51 | 4 |
| Pays de la Loire | 0.96 | 57 | 1 | 0.93 | 57 | 1 | 0.92 | 57 | 1 |
| Bretagne | 0.92 | 53 | 1 | 0.86 | 53 | 1 | 0.83 | 53 | 1 |
| Poitou-Charente | 0.81 | 46 | 3 | 0.71 | 47 | 2 | 0.66 | 48 | 2 |
| Aquitaine | 0.87 | 52 | 4 | 0.79 | 53 | 3 | 0.75 | 53 | 3 |
| Midi-Pyrénées | 0.91 | 54 | 3 | 0.84 | 54 | 3 | 0.81 | 54 | 2 |
| Limousin | 0.67 | 54 | 8 | 0.55 | 54 | 7 | 0.49 | 54 | 6 |
| Rhône-Alpes | 0.90 | 55 | 2 | 0.83 | 55 | 2 | 0.79 | 55 | 2 |
| Auvergne | 0.74 | 51 | 4 | 0.62 | 52 | 3 | 0.56 | 52 | 3 |
| Languedoc-Roussillon | 0.93 | 55 | 2 | 0.87 | 55 | 2 | 0.84 | 55 | 2 |
| PACA | 0.88 | 62 | 8 | 0.80 | 61 | 7 | 0.76 | 61 | 7 |
| Corse | 0.32 | 52 | 8 | 0.24 | 52 | 5 | 0.20 | 52 | 4 |

Source : ENTD 2007-2008.

TABLE A.17 – *Estimations pseudo-*EBLUP *standard et modifiées de la durée totale des déplacements par personne un jour de semaine. L'estimation pseudo-*EBLUP *de type compromis utilise une constante de réglage (tuning constant) obtenue à partir des variables nombre de déplacements, distance totale parcourue et durée totale des déplacements. L'estimation pseudo-*EBLUP *de type transfert utilise une constante de réglage obtenue à partir des variables nombre de déplacements et distance totale parcourue.*

| | Pseudo-EBLUP | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Standard | | | Compromis | | | Transfert | | |
| | $\hat{\gamma}_{iw}$ | est. | var. | $\hat{\Gamma}_{iw}$ | est. | var. | $\hat{\Gamma}_{iw}$ | est. | var. |
| Ile-de-France | 0.97 | 72 | 1 | 0.94 | 72 | 1 | 0.93 | 71 | 1 |
| Champagne-Ardenne | 0.64 | 49 | 3 | 0.52 | 50 | 3 | 0.46 | 50 | 3 |
| Picardie | 0.70 | 57 | 6 | 0.58 | 57 | 5 | 0.52 | 57 | 5 |
| Haute-Normandie | 0.56 | 54 | 6 | 0.44 | 54 | 5 | 0.38 | 54 | 4 |
| Centre | 0.70 | 48 | 4 | 0.58 | 49 | 4 | 0.52 | 50 | 3 |
| Basse-Normandie | 0.59 | 46 | 4 | 0.48 | 47 | 3 | 0.42 | 48 | 3 |
| Bourgogne | 0.64 | 50 | 5 | 0.52 | 51 | 4 | 0.46 | 51 | 4 |
| Nord | 0.79 | 52 | 4 | 0.68 | 53 | 3 | 0.63 | 53 | 3 |
| Lorraine | 0.72 | 53 | 6 | 0.60 | 53 | 5 | 0.54 | 53 | 4 |
| Alsace | 0.63 | 59 | 4 | 0.51 | 58 | 3 | 0.45 | 58 | 3 |
| Franche-Comté | 0.62 | 54 | 9 | 0.50 | 54 | 7 | 0.44 | 54 | 6 |
| Pays de la Loire | 0.91 | 55 | 2 | 0.84 | 55 | 2 | 0.80 | 55 | 1 |
| Bretagne | 0.86 | 55 | 2 | 0.77 | 55 | 2 | 0.72 | 55 | 2 |
| Poitou-Charente | 0.70 | 49 | 6 | 0.59 | 49 | 5 | 0.53 | 50 | 5 |
| Aquitaine | 0.79 | 50 | 4 | 0.69 | 50 | 3 | 0.63 | 51 | 3 |
| Midi-Pyrénées | 0.84 | 53 | 3 | 0.75 | 53 | 3 | 0.70 | 53 | 3 |
| Limousin | 0.53 | 52 | 8 | 0.42 | 52 | 6 | 0.36 | 52 | 5 |
| Rhône-Alpes | 0.83 | 55 | 3 | 0.73 | 55 | 2 | 0.68 | 55 | 2 |
| Auvergne | 0.59 | 50 | 5 | 0.47 | 51 | 4 | 0.41 | 51 | 3 |
| Languedoc-Roussillon | 0.87 | 54 | 3 | 0.79 | 54 | 2 | 0.75 | 54 | 2 |
| PACA | 0.80 | 64 | 15 | 0.70 | 63 | 12 | 0.65 | 62 | 11 |
| Corse | 0.19 | 52 | 4 | 0.14 | 52 | 3 | 0.11 | 52 | 3 |

Source : ENTD 2007-2008.

TABLE A.18 – *Estimations par scission de la durée totale des déplacements par personne un jour de semaine. Les constantes de réglages proviennent respectivement de l'estimateur* EBLUP *et de l'estimateur pseudo-*EBLUP*. La constante de réglage est dite spécifique lorsque celle-ci dépend uniquement de la durée totale par personne un jour de semaine. Elle est dite unique lorsque celle-ci dépend des variables nombre de déplacements, distance totale parcourue et durée totale des déplacements.*

| | Constantes de réglage | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | tirées de l'EBLUP | | | | tirées du pseudo-EBLUP | | | |
| | spécifique | | unique | | spécifique | | unique | |
| | est. | var. | est. | var. | est. | var. | est. | var. |
| Ile-de-France | 71 | 1 | 71 | 1 | 71 | 1 | 70 | 1 |
| Champagne-Ardenne | 49 | 4 | 49 | 3 | 49 | 4 | 50 | 3 |
| Picardie | 57 | 6 | 57 | 5 | 57 | 6 | 57 | 5 |
| Haute-Normandie | 53 | 4 | 54 | 3 | 54 | 3 | 54 | 2 |
| Centre | 49 | 4 | 50 | 4 | 50 | 4 | 51 | 3 |
| Basse-Normandie | 45 | 4 | 47 | 3 | 46 | 4 | 48 | 3 |
| Bourgogne | 48 | 4 | 49 | 3 | 49 | 3 | 50 | 3 |
| Nord | 53 | 4 | 53 | 3 | 54 | 3 | 54 | 3 |
| Lorraine | 52 | 7 | 53 | 6 | 53 | 6 | 53 | 5 |
| Alsace | 59 | 4 | 58 | 3 | 58 | 4 | 58 | 3 |
| Franche-Comté | 55 | 9 | 55 | 7 | 55 | 8 | 55 | 6 |
| Pays de la Loire | 55 | 2 | 55 | 1 | 54 | 2 | 55 | 1 |
| Bretagne | 55 | 2 | 55 | 2 | 55 | 2 | 55 | 2 |
| Poitou-Charente | 48 | 5 | 48 | 4 | 48 | 4 | 49 | 4 |
| Aquitaine | 50 | 4 | 51 | 4 | 51 | 4 | 51 | 4 |
| Midi-Pyrénées | 52 | 3 | 52 | 3 | 52 | 3 | 52 | 3 |
| Limousin | 53 | 11 | 53 | 8 | 53 | 9 | 53 | 7 |
| Rhône-Alpes | 56 | 3 | 56 | 3 | 56 | 3 | 56 | 3 |
| Auvergne | 50 | 5 | 50 | 4 | 50 | 5 | 51 | 3 |
| Languedoc-Roussillon | 54 | 2 | 53 | 2 | 53 | 2 | 53 | 2 |
| PACA | 64 | 12 | 63 | 11 | 63 | 11 | 62 | 9 |
| Corse | 53 | 9 | 53 | 5 | 53 | 7 | 53 | 3 |

Source : ENTD 2007-2008.

# Bibliographie


Antal, E. & Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association* 106 534 – 543. (Cité en pages 25, 93, 138 et 139.)

Ardilly, P. (2006a). *Les Techniques de Sondage*. Technip. (Cité en page 39.)

Ardilly, P. (2006b). *Panorama des principales méthodes d'estimation sur les petits domaines*. Insee. Document de travail. (Cité en page 42.)

Ardilly, P. & Tillé, Y. (2003). *Exercices corrigés de méthodes de sondage*. Ellipses. (Cité en page 39.)

Ardilly, P. & Tillé, Y. (2005). *Sampling Methods : Exercises and Solutions*. Springer. (Cité en page 39.)

Armoogum, J. & Roux, S. (2012). Mise en perspective des Enquêtes Nationales Transports 1973/74 – 1981/82– 1993/94 – 2007/08. Tech. rep., Ifsttar. (Cité en pages 61, 119, 128 et 131.)

Armoogum, J., Roux, S., Hubert, J.-P., Francois, D., Roumier, B. & Robin, M. (2012). Enquête nationale sur les transports et les déplacements 2007-2008. Tech. rep., Ifsttar. (Cité en page 132.)

Arora, H. R. & Brackstone, G. J. (1977). An investigation of the properties of raking ratio estimator : I. With simple random sampling. *Survey Methodology* 3 62–83. (Cité en page 79.)

Battese, G. E., Harter, R. M. & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* 83 28 – 36. (Cité en pages 50, 82, 83, 90, 92, 96, 98, 115, 116 et 117.)

Brewer, K. R. W. (1963). Ratio estimation in finite populations : some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics* 5 93–105. (Cité en page 38.)

Brewer, K. R. W. (1994). Survey sampling inference : Some past perspectives and present prospects. *Pakistan Journal of Statistics* 10 213–233. (Cité en page 36.)

Brewer, K. R. W. (1999). Design-based or prediction-based inference ? stratified random vs stratified balanced sampling. *International Statistical Review* 67 35–47. (Cité en page 36.)

Brewer, K. R. W., Hanif, M. & Tam, S. M. (1988). How nearly can model-based prediction and design-based estimation be reconciled. *Journal of the American Statistical Association* 83 128–132. (Cité en page 36.)

Cassel, C.-M., Särndal, C.-E. & Wretman, J. H. (1977). *Foundations of inference in survey sampling*. Wiley. (Cité en page 36.)

Chauvet, G. (2011). Introduction à l'estimation sur petits domaines. Notes de cours. Ensai. (Cité en page 42.)

Chauvet, G. & Goga, C. (2012). Redresser un échantillon ... mais pas trop. Notes de cours. 44es Journées de Statistique, Bruxelles. (Cité en page 58.)

Cochran, W. G. (1977). *Sampling Techniques*. New York : Wiley. (Cité en page 36.)

Deming, W. E. & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics* 11 427–444. (Cité en pages 40 et 58.)

Deville, J.-C. (1998). *Pour essayer d'en finir avec l'individu* Kish. Insee. Document de Travail. (Cité en pages 137 et 140.)

Deville, J.-C. & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87 376 – 282. (Cité en pages 21, 40, 41, 44, 57, 58, 59, 60, 72, 75, 76 et 145.)

Deville, J.-C., Särndal, C.-E. & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* 88 1013–1020. (Cité en pages 44 et 60.)

Deville, J.-C. & Tillé, Y. (2004). Efficient balanced sampling : The cube method. *Biometrika* 91 893–912. (Cité en page 36.)

Drew, D., Singh, M. P. & Choudhdry, G. H. (1982). Evaluation of small area estimation techniques for the Canadian labour force survey. *Survey Methodology* 8 17–47. (Cité en page 46.)

Fay, R. E. & Herriot, R. A. (1979). Estimates of income for small places : An application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74 269–277. (Cité en pages 48, 49, 50 et 72.)

Fuller, W. A. & Battese, G. E. (1973). Transformation for estimation of linear models with nested error structure. *Journal of the American Statistical Association* 68 626–632. (Cité en pages 72, 83 et 98.)

Gonzalez, M. E. (1973). Use and evaluation of synthetic estimates. In *Proceedings of the Social Statistics Section*. American Statistical Society, 33 – 36. (Cité en page 72.)

Gonzalez, M. E. & Wakesberg, J. (1973). Estimation of the error of synthetic estimates Paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria. (Cité en pages 46, 148, 154 et 160.)

Guggemos, F. & Tillé, Y. (2009). Comparison of two blup estimators under a mixed model for small domain estimation. Unpublished paper. Université de Neuchâtel. (Cité en pages 51 et 95.)

Hansen, M. H., Madow, W. G. & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association* 78. (Cité en page 36.)

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrika* 31 423 – 447. (Cité en pages 30, 50, 51, 52, 53, 82, 83, 95, 96, 98 et 134.)

Henderson, C. R., Kempthorne, O., Searle, S. R. & von Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics* 15 192–218. (Cité en pages 51 et 52.)

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47 663 – 685. (Cité en pages 37, 38, 41, 43, 44, 60, 72, 75, 133 et 136.)

Ireland, C. T. & Kullback, S. (1968). Contingency tables with given marginals. *Biometrika* 55 179–188. (Cité en page 79.)

James, W. & Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley : University of California Press, 361 – 379. (Cité en page 47.)

Kish, L. (1965). *Survey Sampling*. New York : Wiley. (Cité en page 36.)

Lahiri, P., Chambers, R., Rao, J. N. K. & Pfeffermann, D. (2011). European Course on Advanced Statistics on Small Area Statistics. Notes de cours. Universität Trier. (Cité en page 42.)

Le Guennec, J. (2012). Application de méthodes “petits domaines” à des estimations régionales dans l’Enquête Nationale sur les Transports et les Déplacements 2007-2008. In *Actes des Journées de Méthodologie Statistique 2012*. (Cité en pages 64 et 133.)

Le Guennec, J. & Sautory, O. (2002). Application du calage généralisé à la correction de la non-réponse : une expérimentation. In *Actes des Journées de Méthodologie Statistique 2002*. (Cité en page 64.)

Lemel, Y. (1976). Une généralisation de la méthode du quotient pour le redressement des enquêtes par sondages. *Annales de l'Insee* 273–281. (Cité en pages 40 et 58.)

Little, R. (2004). To model or not to model ? competing modes of inference for finite population sampling. *Journal of the American Statistical Association* 99 546–556. (Cité en page 36.)

Madre, J.-L. (1979). *Ajustement et extrapolation de tableaux statistiques*. Ph.D. thesis, Université Pierre et Marie Curie. (Cité en pages 40 et 58.)

Madre, J.-L. (1980). Méthode d'ajustement d'un tableau à des marges. *Les cahiers de l'Analyse des données* 87 – 99. (Cité en pages 40 et 58.)

Molina, I. & Rao, J. N. K. (2012). Small area estimation with examples in R. (Cité en pages 21, 40 et 42.)

Nedyalkova, D. & Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika* 95 521–537. (Cité en page 36.)

Perec, G. (1966). *Quel petit vélo à guidon chromé au fond de la cour ?* Denoël. (Cité en page 171.)

Perec, G. (1989). *La Disparition*. Paris : Gallimard. Collection L'Imaginaire. (Cité en page 42.)

Prasad, N. G. N. & Rao, J. N. K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association* 85 163–171. (Cité en pages 50, 72, 82, 83, 96 et 98.)

Prasad, N. G. N. & Rao, J. N. K. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology* 25 67–72. (Cité en pages 53 et 72.)

Purcell, N. J. & Kish, L. (1979). Estimates for small domains. *Biometrics* 35 365–384. (Cité en page 47.)

Randrianasolo, T., Roux, S. & Armoogum, J. (2010). étude exploratoire sur la connaissance de la mobilité à l'échelle régionale – Analyse de l'Enquête Nationale sur les Transports et les Déplacements 2007–2008. Tech. rep., Département Économie et Sociologie des Transports – Inrets. (Cité en pages 19, 119, 120, 121, 123, 124, 125 et 127.)

Randrianasolo, T. & Tillé, Y. (2013). Small area estimation by splitting the sampling weights. *Electronic Journal of Statistics* 7 1835–1855. (Cité en pages 29 et 71.)

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. New York : John Wiley. (Cité en pages 83 et 98.)

Rao, J. N. K. (2003). *Small Area Estimation*. New-York : Wiley. (Cité en pages 42, 50, 55, 71, 72, 81, 82, 83, 84, 91, 96, 97, 98, 109 et 115.)

Roux, S. (2012). *Transition de la motorisation en France au XX$^{e}$ siècle*. Ph.D. thesis, Université Paris-Sorbonne. (Cité en pages 21 et 62.)

Roux, S. & Armoogum, J. (2008). Correction de la non-réponse dans l'Enquête Nationale sur les Transports et les Déplacements 2007-2008. Tech. rep., Département Économie et Sociologie des Transports – Inrets. (Cité en pages 19, 58, 61, 62, 119, 120, 121, 123, 124, 125, 127, 128 et 131.)

Roux, S. & Armoogum, J. (2010). Redressement de l'Enquête Nationale sur les Transports et les Déplacements 2007-2008. Tech. rep., Département Économie et Sociologie des Transports – Inrets. (Cité en pages 58, 61, 62, 119, 124, 125, 127, 128 et 131.)

Roux, S. & Armoogum, J. (2011). Calibration strategies to correct nonresponse in a national travel survey. *Transportation Research Records : Journal of the Transportation Research Board* 2246 1 – 7. (Cité en pages 61 et 62.)

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* 57 377–387. (Cité en pages 38, 51, 52 et 95.)

Royall, R. M. (1971). Linear regression models in finite population sampling theory *In* Godambe, V. P. et Sprott, D. A. Éds., *Foundations of Statistical Inference*, Toronto, Montréal. (Cité en page 38.)

Royall, R. M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association* 71 657–664. (Cité en page 38.)

Royall, R. M. (1988). The prediction approach to sampling theory. In *Sampling*, Krishnaiah, P. R. & Rao, C. R., vol. 6 of *Handbook of Statistics*. Amsterdam, Holland : Elsevier, 399–413. (Cité en page 36.)

Särndal, C.-E. (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics* 5 27–52. (Cité en page 36.)

Särndal, C.-E., Swensson, B. & Wretman, J. H. (1992). *Model Assisted Survey Sampling*. Springer. (Cité en pages 36, 39, 42 et 72.)

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* 5 119–127. (Cité en page 37.)

Smith, T. M. F. (1994). Sample surveys 1975–1990 ; an age of reconciliation (with discussion) ? *International Statistical Review* 62 5–34. (Cité en page 36.)

Stephan, F. F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *The Annals of Mathematical Statistics* 13 166–178. (Cité en pages 40 et 58.)

Tillé, Y. (1992). *Utilisation a posteriori d'informations auxiliaires en théorie des sondages sans référence à un modèle*. Ph.D. thesis, Université Libre de Bruxelles. (Cité en pages 38, 39 et 57.)

Tillé, Y. (2001). *Théorie des sondages : Échantillonnage et estimation en populations finies*. Dunod. (Cité en pages 28, 31, 38, 40, 44 et 60.)

Ugarte, M., Militino, A. & Goicoa, T. (2009). Benchmarked estimates in small areas using linear mixed models with restrictions. *Test* 18 342–364. (Cité en page 73.)

Valliant, R., Dorfman, A. H. & Royall, R. M. (2000). *Finite Population Sampling and Inference : A prediction Approach*. Wiley. (Cité en page 38.)

Wilms, L. (2000). L'échantillon-Maître 99 et Application au tirage des unités primaires par la macro "Cube". In *Actes des Journées de Méthodologie Statistique 2000*. (Cité en page 125.)

Winkler, B. (2011). *Blagues mathématiques et autres curiosités*. Ellipses. (Cité en page 7.)

Yates, F. & Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society. Series B. Methodological* 15 235–261. (Cité en page 37.)

You, Y. & Rao, J. N. K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics* 30 431–439. (Cité en pages 30, 50, 53, 54, 73, 82, 83, 84, 90, 95, 96, 98, 99, 100 et 115.)

You, Y. & Rao, J. N. K. (2003). Pseudo hierarchical bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference* 111 197 – 208. (Cité en page 73.)

Ce document a été préparé à l'aide de l'éditeur de texte TEXShop et du logiciel de composition typographique LATEX 2ε.
Les *fleurs et ornements rhétoriques* utilisés dans ce document sont largement inspirés du livre *Quel petit vélo à guidon chromé au fond de la cour ?* de Perec (1966).

## Inférence basée sur le plan pour l'estimation de petits domaines

**Résumé** La forte demande de résultats à un niveau géographique fin, notamment à partir d'enquêtes nationales, a mis en évidence la fragilité des estimations sur petits domaines. Cette thèse propose d'y remédier avec des méthodes spécifiques basées sur le plan de sondage. Celles-ci reposent sur la construction de nouvelles pondérations pour chaque unité statistique. La première méthode consiste à optimiser le redressement du sous-échantillon d'une enquête inclus dans un domaine. La deuxième repose sur la construction de poids dépendant à la fois des unités statistiques et des domaines. Elle consiste à scinder les poids de sondage de l'estimateur global tout en respectant deux contraintes : 1/ la somme des estimations sur toute partition en domaines est égale à l'estimation globale ; 2/ le système de pondération pour un domaine particulier satisfait les propriétés de calage sur les variables auxiliaires connues pour le domaine. L'estimateur par scission ainsi obtenu se comporte de manière *quasi* analogue au célèbre estimateur BLUP (meilleur prédicteur linéaire sans biais). La troisième méthode propose une réécriture de l'estimateur BLUP sous la forme d'un estimateur linéaire homogène, en adoptant une approche basée sur le plan de sondage, bien que l'estimateur dépende d'un modèle. De nouveaux estimateurs BLUP modifiés sont obtenus. Leur précision, estimée par simulation avec application sur des données réelles, est assez proche de celle de l'estimateur BLUP standard. Les méthodes développées dans cette thèse sont ensuite appliquées à l'estimation d'indicateurs de la mobilité locale à partir de l'Enquête Nationale sur les Transports et les Déplacements 2007-2008. Lorsque la taille d'un domaine est faible dans l'échantillon, les estimations obtenues avec la première méthode perdent en précision, alors que la précision reste satisfaisante pour les deux autres méthodes.

**Mots-clés** sondage, estimation sur petits domaines, inférence basée sur le plan de sondage, poids.

## Design-based inference for small area estimation

**Abstract** The strong demand for results at a detailed geographic level, particularly from national surveys, has raised the problem of the fragility of estimates for small areas. This thesis addresses this issue with specific methods based on the sample design. These ones consist of building new weights for each statistical unit. The first method consists of optimizing the re-weighting of a subsample survey included in an area. The second one is based on the construction of weights that depend on the statistical units as well as the areas. It consists of splitting the sampling weights of the overall estimator while satisfying two constraints : 1/ the sum of the estimates on every partition into areas is equal to the overall estimate ; 2/ the system of weights for a given area satisfies calibration properties on known auxiliary variables at the level of the area. The split estimator thus obtained behaves almost similarly as the well-known BLUP (best linear unbiased predictor) estimator. The third method proposes a rewriting of the BLUP estimator, although model-based, in the form of a homogenous linear estimator from a design-based approach. New modified BLUP estimators are obtained. Their precision, estimated by simulation with an application to real data, is quite close to that of the standard BLUP estimator. Then, the methods developed in this thesis are applied to the estimation of local mobility indicators from the 2007-2008 French National Travel Survey. When the size of an area is small in the sample, the estimates obtained with the first method are not precise enough whereas the precision remains satisfactory for the two other methods.

**Keywords** survey sampling, small area estimation, design-based inference, weights.