ISSN 1680-8789

SERIE

ÉTUDES STATISTIQUES 97

Désagrégation des données pour les enquêtes auprès des ménages

Utilisation de méthodes d'estimation sur petits domaines

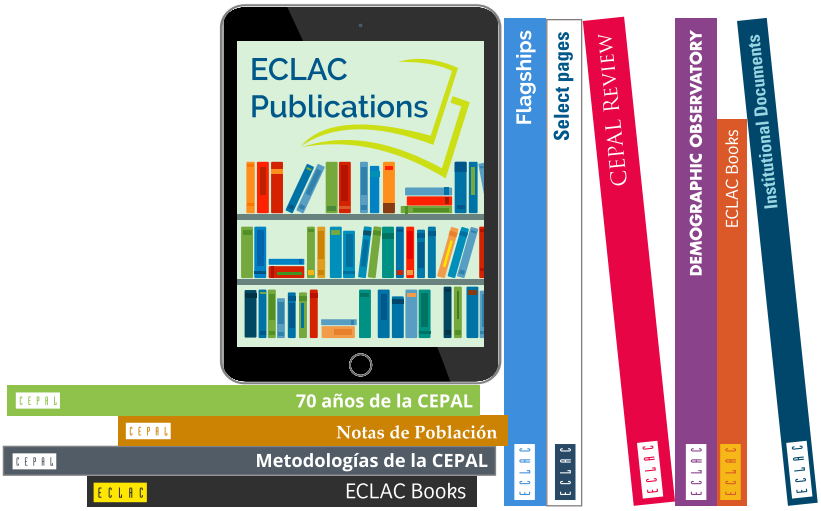
Isabel Molina

NATIONS UNIES

CEPALC

United Nations | DESA Statistics

Thank you for your interest in this ECLAC publication



Please register if you would like to receive information on our editorial products and activities. When you register, you may specify your particular areas of interest and you will gain access to our products in other formats.

Register

www.cepal.org/en/publications

www.instagram.com/publicacionesdelacepal

www.facebook.com/publicacionesdelacepal

www.issuu.com/publicacionescepal/stacks

www.cepal.org/es/publicaciones/apps

UNITED NATIONS

ECLAC

SERIE

Études statistiques	97

Désagrégation des données pour les enquêtes auprès des ménages

Utilisation de méthodes d'estimation sur petits domaines

Isabel Molina

NATIONS UNIES
CEPALC

United Nations | DESA Statistics

La version espagnole de ce document a été produite par Isabel Molina, chercheur au Département de Statistiques, Carlos III University à Madrid, en collaboration avec Andrés Gutiérrez, expert en statistiques sociales, et Álvaro Fuentes, statisticien, tous deux à l'Unité de Statistiques Sociales de la Division des Statistiques de la Commission Economique pour l'Amérique Latine et les Caraïbes (ECLAC), dans le cadre de la dixième tranche du *United Nations Development Account*. Cette traduction en français a été rendue possible grâce au financement de la Division de statistique des Nations Unies, afin que le texte soit utilisé comme document de référence pour les activités de renforcement réalisées dans la treizième tranche du *United Nations Development Account*. .

Les opinions exprimées dans ce document, une traduction non officielle dont l'original n'a pas été soumis à une révision éditoriale formelle, sont de la responsabilité exclusive des auteurs et peuvent ne pas coïncider avec celles de l'Organisation ou des pays qu'elle représente.

Publication des Nations Unies
ISSN: 1680-8789 (version électronique)
ISSN: 1994-7364 (version papier)
LC/TS.2018/82/REV.1
Distribution: L

Imprimé aux Nations Unies, Santiago
S.22-00807

Cette publication doit être citée comme suit: I. Molina, « Désagrégation des données pour les enquêtes auprès des ménages: utilisation de méthodes d'estimation sur petits domaines », *série Études statistiques*, N°97 (LC/TS.2018/82/REV.1), Santiago, Commission économique pour l'Amérique latine et les Caraïbes (CEPALC), 2022.

Index

Tableaux

Figures

Résumé

Les enquêtes auprès des ménages sont largement utilisées en tant qu'outils destinés à produire de l'information sur les conditions socio-économiques et le bien-être des ménages. Mais la précision des estimations obtenues à partir de ces enquêtes diminue de manière significative quand on cherche à les produire sur des groupes de population qui sont des sous-ensembles que le plan de sondage de l'enquête n'a pas pris en compte a priori. Dans ce contexte, il est possible d'utiliser des procédures d'estimation combinant l'information de l'enquête avec de l'information auxiliaire disponible pour la population, par exemple issue de recensements ou de fichiers administratifs.

Ce document constitue un guide méthodologique présentant l'utilisation combinée de techniques statistiques d'enquêtes et de modèles probabilistes destinée à produire des statistiques désagrégées pour des groupes d'intérêt, connue sous le nom d'estimation sur petits domaines (en anglais : SAE, comme small area estimation[1]).

Une description du problème posé par la désagrégation dans le cas où l'on dispose de trop peu de données précède la présentation de trois méthodes destinées à proposer des solutions pour y répondre. On passe d'abord en revue les estimateurs directs (s'appuyant de manière directe sur les enquêtes), qui ont l'avantage d'être non biaisés, mais avec une précision faible qaund on procède à de la désagrégation. Dans un second temps, on présente certains estimateurs appelés estimateurs indirects, dont la forme fonctionnelle est similaire à celle des estimateurs directs, mais qui s'appuient sur une information auxilaire au niveau de la population pour améliorer la precision. Ensuite, les modèles probabilistes sont introduits afin d'améliorer les propriétés statistiques des estimateurs. La modélisation peut être faite à deux niveaux : au niveau individuel (ménages ou personnes) ou au niveau des catégories auxquelles on s'intéresse (sous-groupes sur lesquels on cherche à produire de l'information). La discussion théorique est complétée par des illustrations et des exemples qui s'appuient sur le logiciel statistique R. Enfin, une application pratique est donnée pour certaines des méthodes qui ont été présentées, et des conclusions sont tirées sur la faisabilité de l'utilisation de ces méthodes.

[1] Dans la suite de ce doument, on conservera les acronymes anglais.

Introduction

Les enquêtes auprès des ménages fournissent une information fondamentale pour mesurer les conditions de vie de la population d'un pays, et constituent un outil essentiel pour définir et suivre des politiques publiques dans de nombreux domaines. Elles permettent de produire des estimations fiables et sans biais au niveau national, et pour des niveaux de désagrégation qui ont été pris en compte au moment de la conception de l'enquête.

Il existe une demande de plus en plus importante d'information pour des groupes de populations spécifiques, et pour des petites zones géographiques. Par exemple, le cadre général des indicateurs destinés à suivre le Développement Durable indique que l'information devrait être désagrégée non seulement au niveau géographique (pour des découpages comme les régions, les municipalités, les districts), mais également pour des groupes de revenu, par sexe, par âge, par race, par origine ethnique, par statut concernant l'immigration, ou par statut relativement au handicap. Cependant, la fiabilité des inférences tirées des indicateurs diminue quand la taille d'échantillon concernée diminue, ce qui conduit généralement à l'impossibilité de produire les estimations souhaitées à ces niveaux avec une précision suffisante.

C'est pourquoi il y a eu, dans la dernière décennie, des avancées significatives concernant le concept de désagrégation des données, à savoir la compilation sous forme agrégée et résumée d'informations numériques collectées à partir de différentes sources ou mesurées par des variables multiples, éventuellement sur différentes unités d'observation. Le but recherché est de fournir à la société des estimations ayant de bonnes propriétés statistiques, et qui peuvent être utilisées pour extraire de l'information, voire formuler des politiques publiques, pour les différents sous-groupes d'intérêt.

Ce document est un guide présentant la désagrégation des données statistiques concernant les conditions de vie des personnes, soit au niveau géographique (au niveau régional), soit pour des sous-groupes de populations. Le chaptitre I commence par décrire le problème de la désagrégation des données statistiques (section I.A); il décrit dans quels contextes ce problème est posé, et définit les termes et concepts utilisés de manière courante et qui seront donc repris dans la suite de ce document.

La section I.B pose la question du niveau auquel il est approprié de désagréger les données statistiques, étant donné que, en raison de la diminution des tailles d'échantillon, plus les estimations directes sont désagrégées plus les erreurs d'échantillonage augmentent, rendant les estimateurs trop instables et de ce fait peu fiables. Par exemple, on peut considérer une population divisée successivement à différents niveaux; l'Espagne, par exemple, est divisée en communautés autonomes, elles-mêmes divisées en provinces; celles-ci sont divisées en régions, et enfin les régions sont divisées en municipalités. Au niveau de l'ensemble de l'Union Européenne, la nomenclature commune NUTS (Nomenclature of Territorial Units for Statistics) est utilisée et les pays (NUTS 0) sont divisés en régions appelées NUTS 1, NUTS 2, etc. La section I.B fournit des indications sur le niveau maximum de désagrégation des estimateurs directs, à partir duquel il faut envisager d'utiliser des estimateurs indirects. Ces derniers sont beaucoup plus fiables car ils s'appuient sur différentes sources de données pour "emprunter" de l'information sur l'ensemble des zones. Elle aborde également la question des limites de l'utilisation des estimateurs indirects, car il est conseillé de se prémunir contre leurs biais potentiels. C'est pourquoi des recommandations sont faites sur les cas où il est plus prudent de ne procéder à aucune estimation. Dans tous les cas, la conception de l'enquête pourrait être revue afin d'avoir une meilleure couverture des domaines pour lesquels on souhaite produire des estimations. Il faut aussi se rappeler que, au niveau local, l'information ou le savoir détenus par les communautés locales pourraient être en contradiction avec les estimations proposées. C'est pourquoi il est essentiel de déterminer jusqu'à quel niveau il est possible de désagréger les estimations, afin qu'elles soient de qualité suffisante, réalistes et ne s'écartant pas trop de ce qui est connu au niveau local. Enfin, la section I.C passe en revue différentes méthodologies d'estimations indirectes qui permettent de s'affranchir des limites des estimations directes pour la désagrégation. Plus précisément, elle présente les estimateurs indirects basiques, qui comprennent les estimateurs synthétiques et composites, et les estimateurs basés sur une modélisation, qui sont peut-être les plus usités pour produire des estimations fiables à des niveaux très désagrégés. Les estimateurs "assistés par un modèle", qui utilisent un modèle mais ne requièrent pas une validité de l'ajustement pour maintenir leur caractère sans biais sont quant à eux présentés dans le chapitre III avec les méthodes directes, car ils ont de bonnes propriétés théoriques pour les grandes tailles d'échantillon.

Le chapitre II présente différents indicateurs concernant la "qualité de la vie" individuelle; en particulier la mesure de la pauvreté et des inégalités. Une famille d'indicateurs de mesure de pauvreté, appelée famille FGT, est définie de manière détaillée et sera utilisée pour illustrer les différentes procédures dans les chapitres suivants. Une description de chaque procédure explique comment elle est appliquée à l'estimation des indicateurs de cette famille, et, pour une part d'entre eux, des exemples sont donnés à partir du package R sae (Molina et Marhuenda, 2015), que le lecteur peut récupérer.

Le chapitre III donne une description détaillée des estimateurs directs habituels. Les estimateurs directs basiques comme l'estimateur de Horvitz-Thompson et celui de Hájek sont inclus (section III.A), et également les estimateurs assistés par un modèle; plus précisément, les estimateurs par la régression généralisée et les estimateurs par calage (section III.B), conjointement avec des estimateurs de leurs erreurs d'échantillonnage. La mise en ouvre de ces estimateurs sous R est illustrée au moyen de deux exemples.

Le chapitre IV passe en revue certains estimateurs indirects de base, comme l'estimateur synthétique post-stratifié (section IV.A), l'estimateur synthétique par la régression au niveau domaine (section IV.B) ou au niveau individu (section IV.C), et les estimateurs composites (section IV.D). Ces estimateurs sont présentés ici essentiellement parce qu'ils donnent un bon aperçu des idées qui sous-tendent les méthodes plus sophistiquées présentées plus loin au chapitre V. A nouveau, deux exemples de calcul des estimateurs synthétique et composite post-stratifés sont présentés.

Les méthodes basées sur un modèle présentées dans le chapitre V sont, de manière significative, plus réalistes que les méthodes indirectes basiques, et sont plus appropriées pour les applications

"réelles", car elles conduisent à des estimations potentiellement moins biaisées. Ces méthodes basées sur un modèle comprennent la catégorie la plus populaire des estimateurs basés sur un modèle au niveau domaine (section V.A), et celle des estimateurs basés sur un modèle au niveau individuel (section V.B). Trois exemples présentent comment obtenir ces estimateurs en R. La section V.C s'intéresse à la méthode ELL, utilisée par la Banque Mondiale pour estimer des indicateurs de pauvreté et/ou d'inégalité, méthode qui peut être rattachée à la section précédente puisque dans son principe elle considère le modèle au niveau individuel. Néanmoins, comme nous le verrons, cette méthode est essentiellement synthétique, et pourrait de ce fait être rattachée au chapitre IV qui traite des estimateurs synthétiques. Il y a aussi une description de la méthode EB (section V.D), qui estime des indicateurs généraux de la même façon que la méthode ELL, mais l'améliore en considérant qu'il existe une hétérogénéité entre les domaines et, de ce fait, produit des estimations plus précises. La procédure HB dans la section V.E produit des estimations très proches de celles de la méthode EB, mais avec des temps de calcul plus faibles pour des populations nombreuses, en particulier quand on veut produire des mesures de l'erreur (erreur quadratique moyenne) pour ces estimations. Enfin, la section V.F présente des méthodes spécifiques pour estimer des indicateurs qui sont des proportions, ou des moyennes de variables binaires. Bien qu'il soit en principe possible d'utiliser d'autres méthodes pour estimer ces indicateurs, par exemple celles des sections V.A ou V.B, celles-ci pourraient conduire à des estimations se situant en-dehors de l'intervalle attendu pour le ratio. Dans d'autres cas, les estimations obtenues par différentes méthodes ne sont pas très différentes.

Par ailleurs, certaines des méthodes présentées sont applicables seulement à des indicateurs linéaires, c'est-à-dire qui sont additifs relativement aux valeurs d'intérêt des unités du domaine, comme des moyennes ou des totaux. D'autres méthodes, comme les méthodes basées sur un modèle au niveau individuel ELL, EB et HB des sections V.C, V.D et V.E, sont conçues de manière à pouvoir estimer des indicateurs généraux définis comme des fonctions de valeurs d'une variable continue (par exemple le revenu) pour les unités du domaine; valeurs pour lesquelles on postule l'existence d'un modèle. Les méthodes basées sur des modèles au niveau domaine sont, en principe, utilisables pour de nombreux types d'indicateurs, tant que les hypothèses "nécessaires" sont vraies, mais dans la pratique il est difficile de vérifier de telles hypothèses (telles que le caractère sans biais des estimateurs directs) pour les indicateurs non linéaires. C'est pourquoi, en principe, elles sont plus adaptées à l'estimation de moyennes et de totaux. Dans tous les cas, après la présentation de chaque méthode, il y a un résumé qui indique à quels indicateurs elle peut être appliquée, les conditions nécessaires pour les données autres que la variable d'intérêt obtenue via l'enquête, et les avantages et inconvénients de chaque méthode comparée aux autres méthodes utilisables pour le même type d'indicateur.

Il faut noter qu'il n'est pas possible de donner ici une description détaillée de l'ensemble des méthodes existantes, faute de place. Cette description est donc fournie pour les méthodes les plus fréquemment utilisées et ayant les bonnes propriétés. Elle peut aider à la compréhension de méthodes plus complexes. Certaines des extensions des méthodes les plus utilisées sont citées, renvoyant le lecteur à la bibliograhie pour plus d'informations. Les méthodes dont les propriétés théoriques restent inconnues ne sont pas non plus incluses dans ce document, même si elles paraissent prometteuses. De même pour les détails de procédures qui demandent des formulations mathématiques excessivement complexes, comme l'estimation de l'erreur quadratique moyenne des estimateurs au chapitre IV et à la section V.B. Dans les deux cas, le lecteur est renvoyé à la bibliographie où des développements peuvent être trouvés. Dans le cas des estimateurs indirects de base du chapitre IV, une autre raison pour ne pas inclure ces développements est qu'il n'existe pas d'estimateurs connus et fiables de leur erreur quadratique moyenne qui seraient différents selon les domaines. Il existe des estimateurs qui sont très instables mais différents selon les domaines, ou stables mais les mêmes pour chaque domaine, et pas "les deux en même temps". Ce problème reste donc un problème sans solution.

Une clarification est nécessaire concernant l'approche utilisée pour évaluer la qualité de l'estimateur. Il existe trois approches alternatives pour évaluer les propriétés d'un estimateur, mais souvent chaque type d'estimateur est évalué avec un seul type de mesure, lié à l'approche "naturelle" qui sous-tend estimateur. Les estimateurs directs et indirects de base sont évalués au regard de la distribution des échantillons possibles avec le plan de sondage utilisé. Dans ce cas, les valeurs des variables d'intérêt des unités de la population sont considérées comme fixées, et seule la composition de l'échantillon est variable (résultant d'un processus aléatoire). Un bon estimateur est donc celui qui a de bonnes performances en moyenne pour l'ensemble des échantillons possibles, les valeurs des variables dans la popuation étant fixées.

Pour leur part, les méthodes d'estimation basées sur un modèle sont évaluées au regard de la distribution générée par le modèle considéré, conditonnellement à l'échantillon observé. En d'autres termes, les valeurs de la variable d'intérêt des individus de la population sont considérées comme aléatoires, et générées par un modèle appelé modèle de superpopulation. Selon cette approche, le recensement de cette variable est une réalisation possible du vecteur aléatoire qui suit le modèle (ou distribution de probabilité). Les estimateurs sont alors évalués au regard de tous les recensements possibles qui pourraient être générés par le modèle. En d'autres termes, un bon estimateur doit avoir de bonnes performances en moyenne sur l'ensemble (infini) des recensements possibles des valeurs de la variable d'intérêt générés par le modèle, tout en laissant la composition de l'échantillon constante (mais les valeurs de la variable d'intérêt sur l'échantillon varient, puisque provenant d'un des recensements possibles).

Enfin, les méthodes bayésiennes, telles que la méthode HB dans la section V.E, sont évaluées conditionnellement aux observations de la variable d'intérêt dans l'échantillon (distribution a posteriori). En d'autres termes, un estimateur sera évalué au regard de la distribution de l'indicateur conditionnellement aux données disponibles, plutôt que "moyenné" sur l'ensemble des valeurs possibles des données.

Il ny a pas de consensus sur ce que serait une approche optimale pour évaluer les méthodes d'estimation sur petits domaines. L'approche "sous le plan de sondage" est non paramétrique en ce sens qu'elle ne postule pas l'existence d'un modèle. Ce qui signifie que la mesure de l'erreur proposée selon cette aproche (en général l'erreur quadratique moyenne) prend en compte l'erreur d'estimation parmi l'ensemble des échantillons possibles, sans qu'il y ait besoin de vérifier les hypothèses d'un modèle. C'est l'approche préférée par les statisticiens des agences gouvernementales. L'approche "sous le modèle" suppose l'existence d'un modèle, mais fixe l'échantillon obtenu, produisant ainsi l'erreur pour l'échantillon en question, plutôt qu'une moyenne sur l'ensemble des échantillons qui pourraient être tirés.

Avec cette approche, la mesure de l'erreur prend en compte l'incertitude au travers de l'ensemble des recensements qui peuvent être générés par le modèle, c'est-à-dire au travers des "réalités possibles" qui pourraient exister, les valeurs variant aussi pour l'échantillon tiré. Enfin, l'approche bayésienne considère les indicateurs comme des réalisations de variables aléatoires qui suivent une distribution, et fournit une mesure de l'erreur sous la forme de valeurs descriptives de la distribution de ces indicateurs, conditionnellement aux valeurs observées sur l'échantillon, plutôt que sur une moyenne obtenue sur un ensemble de valeurs possibles.

Comme indiqué précédemment, chaque méthode d'estimation est en général évaluée sur la base de son approche naturelle. En d'autres termes, les mesures d'erreur qui accompagnent les estimations pour évaluer leur qualité, particulièrement l'erreur quadratique moyenne, sont habituellement calculées selon l'approche utilisée pour l'estimateur. Ceci signifie que les erreurs quadratiques moyennes de différents estimateurs obtenus selon différentes approches ne sont pas directement comparables. Cependant, il est prouvé que, si les hypothèses qui sous-tendent les modèles considérés sont valides,

ces erreurs quadratiques moyennes sont en fait comparables quand on les considère en moyenne sur un grand nombre de domaines de même taille d'échantillon. De plus, les erreurs quadratiques moyennes dans le cas d'estimateurs basés sur un modèle ne sont pas faciles à estimer, et aucun estimateur "acceptable" n'est connu. Mais, d'un autre côté, en procédant à une validation du modèle "a priori" pour vérifier que celui-ci s'ajuste bien aux données disponibles, les estimateurs de l'erreur quadratique moyenne sous le modèle, qui sont relativement stables, peuvent être comparés à l'erreur quadratique moyenne sous le plan de sondage.

I. Le problème de la désagrégation des données (ou estimation sur petits domaines)

A. Description du problème

Les enquêtes officielles menées par les Instituts Nationaux de Statistique, comme par les Bureaux Régionaux de Statistique ou d'autres agences ou institutions situées à un niveau supranational ou international, sont conçues pour produire des données statistiques à un niveau déterminé d'agrégation, que ce soit pour des subdivisions géographiques ou socio-économiques de la population. Par exemple, le module sur les Conditions Socio-économiques (*Módulo de Condiciones Socioeconómicas* (MCS)) de l'enquête nationale mexicaine sur les revenus et les dépenses des ménages (*Encuesta Nacional de Ingresos y Gastos de los Hogares* (ENIGH)) est conçu pour fournir des indicateurs de pauvreté et d'inégalité au niveau national pour l'ensemble des 32 Etats (31 Etats plus Mexico City), désagrégés par zones rurales et urbaines, tous les deux ans. Cependant, il y a une obligation, dans ce pays, de produire aussi des estimations tous les 5 ans au niveau des municipalités. Cette situation est courante pour d'autres pays et d'autres domaines, ce qui signifie que, une fois l'enquête réalisée, avec des tailles d'échantillon déterminées pour produire des estimations fiables à un certain niveau d'agrégation, des demandes de production de données à un niveau plus désagrégé sont souvent formulées. Pour y répondre, nous souhaitons pouvoir utiliser les données de l'enquête sans provoquer des coûts aditionnels qui seraient dus à l'augmentation de la taille de l'échantillon. Mais, dans le cas du Mexique, les sous-échantillons de l'ENIGH pris pour chaque municipalité ne permettent pas d'obtenir des estimations directes fiables dans chacune d'entre elles et, dans la pratique, plus de la moitié des municipalités n'ont pas assez d'observations. Ce problème arrive souvent quand on cherche à produire des données statistiques pour des subdivisions plus petites que celles qui avaient été envisagées au départ.

Pour éviter ce problème (jusqu'à un certain point), des éléments du plan de sondage peuvent être modifiés en amont de la réalisation de l'enquête. Il est par exemple possible d'augmenter les tailles d'échantillon dans les domaines où c'est nécessaire (avec comme conséquence une augmentation du

coût), ou de répartir l'échantillon de manière plus efficace entre les domaines. Mais bien qu'il y ait différentes façons d'améliorer le plan de sondage et d'avoir un nombre suffisant d'observations dans toutes les subdivisions de la population, "le client demande toujours plus que ce qui a été spécifié au moment de la conception du plan de sondage" (Fuller, 1999).

Dans la littérature dédiée, les subdivisions pour lesquelles des données statistiques (ou estimations) sont requises sont souvent appelées « domaines », qu'ils correspondent à des délimitations géographiques ou socio-économiques. Quand on veut estimer un indicateur spécifique pour l'un de ces domaines, on utilise l'appellation "estimateur direct" pour un estimateur qui utilise uniquement les données d'enquête observées sur le domaine. Les estimateurs directs habituels sont sans biais, ou virtuellement sans biais au regard de la distribution d'échantillonnage, c'est-à-dire parmi tous les échantillons qu'on peut tirer dans la population avec le plan de sondage considéré. Mais si l'enquête n'a pas été conçue pour produire des estimations à un niveau aussi détaillé, la taille de l'échantillon "tombant" dans certains domaines peut être trop petite, conduisant à une erreur d'échantillonnage très grande pour les estimateurs directs des indicateurs dans ces domaines. Les domaines pour lesquels cela arrive, relativement à la taille de la population, sont appelés « petits domaines ». De ce fait, ce n'est pas la taille de la population du domaine qui donne la caractéristique "petit domaine", puisque dans beaucoup de cas des domaines de taille importante en termes de population totale (par exemple des Etats aux USA) sont considérés comme « petits domaines » si les estimtions directes n'y sont pas de qualité suffisante. Précisément, le terme « petit domaine » se réfère à des domaines pour lesquels l'estimateur direct d'un indicateur d'intérêt est inefficace, en raison d'un nombre insuffisant d'observations qui y sont obtenues (ou d'enquêtes réalisées). Par exemple, si l'on cherche à produire des estimations des indicateurs de pauvreté et d'inégalité basées sur le Module sur les Conditions Socio-économiques de l'ENIGH mexicaine, les municipalités seront considérées comme des petits domaines, l'enquête n'ayant pas été conçue pour fournir des estimations précises sur celles-ci.

Les estimations produites à un niveau géographique très détaillé sont souvent visualisées sous forme de cartogrammes ou de cartes où les régions sont représentées avec des nuances, ou des couleurs, indiquant différents degrés d'intensité d'un indicateur d'intérêt. Par exemple, la Banque Mondiale produit des cartes détaillant les indicateurs de pauvreté ou d'inégalité pour de nombreux pays (voir par exemple Elbers, Lanjouw et Lanjouw (2003)). Ces cartes, ainsi que les estimations correspondantes, constituent un outil essentiel pour suivre les conditions de vie dans les différentes régions d'un pays, et sont utilisées par les gouvernements et les agences internationales pour planifier des politiques de développement régionales. Il est très recommandé de fournir, à côté des estimations, des indicateurs de qualité de ces estimations (en général l'erreur d'échantillonnage). Comme pour les estimations, ces informations peuvent être également indiquées sur les cartes.

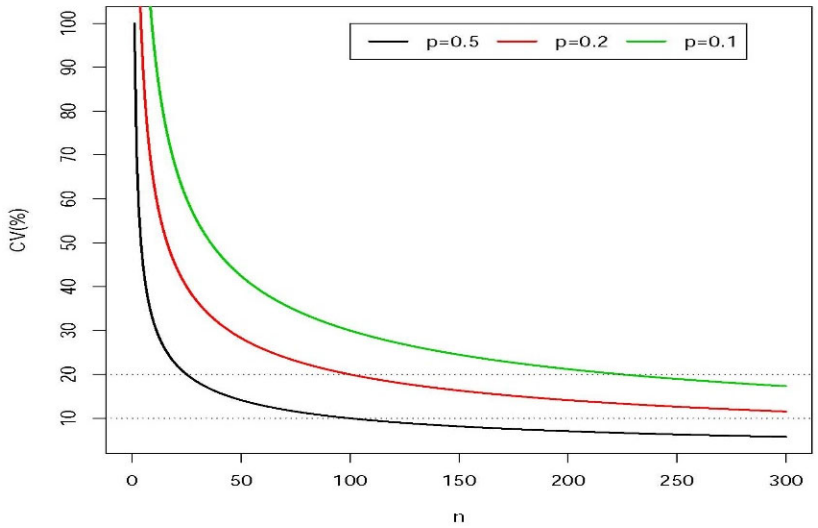
B. Limites de la désagrégation des données statistiques

Bien qu'il n'en existe aucune définition formelle, un domaine est considéré comme "petit", comme indiqué précédemment, quand l'erreur aléatoire de l'estimateur direct relative à l'indicateur d'intérêt n'est pas acceptable. Mais il n'existe pas de limite supérieure universelle pour l'erreur d'échantillonnage, au-dessus de laquelle le domaine à estimer serait considéré comme "petit". Chaque Institut National Statistique ou agence internationale établit sa propre limite pour l'erreur d'échantillonnage relative, ou le coefficient de variation (CV), au-dessus de laquelle une donnée statistique est considérée comme non fiable et donc non publiable. Ce type de données est d'ailleurs parfois publié, avec une indication de son manque de fiabilité. Il n'y a pas non plus de taille d'échantillon en-dessous de laquelle un échantillon est considéré comme petit, car l'erreur d'échantillonnage dépend non seulement de la taille d'échantillon mais également de l'indicateur à estimer, et de l'estimateur utilisé. Par exemple, si on estime la moyenne d'une variable continue (comme le revenu moyen) avec un seuil maximum pour l'erreur

d'échantillonnage, la taille d'échantillon nécessaire est souvent plus petite que celle requise pour l'estimation de la proportion des individus possédant une certaine caractéristique (moyenne d'une variable binaire), en particulier si cette caractéristique est très rare, ou très fréquente, donc si la proportion est proche de 0 ou de 1.

La figure 1 illustre la taille minimum d'échantillon nécessaire pour obtenir une valeur du CV maximum relativement à une proportion échantillonnée à partir d'un sondage alétoire simple. La taille d'échantillon requise varie selon la vraie valeur de la proportion à estimer. Plus précisément, le graphique montre que si la « vraie » proportion est $p = 0.5$, une taille d'échantillon d'environ $n = 25$ est suffisante pour assurer un CV relatif à la proportion calculée sur l'échantillon en dessous de 20%, alors que pour $p = 0.2$ au moins $n = 100$ unités sont nécessaires, et pour $p = 0.1$ plus de $n = 200$ unités sont nécessaires dans l'échantillon. C'est pourquoi il n'est pas possible d'établir une taille minimum d'échantillon dans les domaines permettant de garantir un niveau souhaité d'efficacité pour tout estimateur et/ou tout indicateur cible.

Figure 1
CV de la proportion sur l'échantillon $\hat{p}$ selon la taille de l'échantillon n, pour chaque valeur de la vraie proportion p
(En pourcentage)



Source: Calculs de l'auteur.

Certains indicateurs de pauvreté sont des proportions. Par exemple, l'incidence de la pauvreté, encore appelée taux d'individus en risque de pauvreté, est la proportion d'individus dont le revenu est en-dessous du seuil de pauvreté. Ce seuil est la valeur du revenu (en équivalent net) en-dessous duquel un individu est considéré comme en risque de pauvreté ou d'exclusion. De la même manière, certains types de privations sont mesurés à partir de la proportion d'individus ayant accès à certains services de base, comme la santé, le logement, ou la nourriture. Comme indiqué précédemment, la taille d'échantillon nécessaire pour obtenir des estimations directes de ces indicateurs avec une qualité suffisante est en général plus grande que celle demandée pour estimer des moyennes ou des totaux de variables quantitatives.

Bien qu'il n'y ait pas de limite supérieure « universelle » pour les erreurs d'échantillonnage (et pas de limite inférieure pour les tailles d'échantillon) destinée à décider que des données statistiques sont de qualité acceptable, certains Instituts Nationaux Statistiques considèrent qu'une donnée n'est pas

publiable si son erreur d'échantillonnage relative, ou CV, dépasse 20%. Ainsi, pour ces Instituts, les domaines pour lesquels des estimations directes d'un indicateur d'intérêt ont un CV supérieur à 20% seront considérés comme "petits", pour cet indicateur. Par exemple, les municipalités mexicaines seraient des petits domaines concernant l'estimation des indicateurs de pauvreté à partir de l'ENIGH. Dans ces zones, il serait nécessaire d'augmenter la taille de l'échantillon ou d'utiliser des méthodes "indirectes " pour être en mesure de produire des données statistiques d'une qualité suffisante pour être publiées.

Les méthodes d'estimation "indirectes" prennent en compte la partie de l'échantillon située dans le domaine ou la zone d'intérêt, mais elles utilisent également les données de l'échantillon relatives à d'autres domaines ou zones. Ces estimateurs utilisent de l'information provenant d'autres variables (appelées variables auxiliaires) qui sont liées à la variable d'intérêt. La relation existant entre les variables est considérée comme semblable pour tous les domaines (ou zones), et est représentée par un modèle les liant au travers de paramètres communs. En estimant ces paramètres grâce à l'ensemble des données de l'échantillon (en général cet ensemble est de taille importante), on utilise une quantité importante d'informations, ce qui conduit à des estimateurs plus efficaces (comparés aux estimateurs directs). Ces estimateurs ont tendance à "admettre" un léger biais, en échange d'une amélioration significative de leur efficacité globale évaluée au travers de l'erreur quadratique moyenne.

L'amélioration de l'efficacité apportée par les estimateurs indirects par rapport à celle des estimateurs directs est d'autant plus importante que la taille de l'échantillon dans le domaine est petite. Mais ces méthodes améliorent l'efficacité dans la plupart des domaines, y compris dans certains où la taille d'échantillon peut être grande. Il se touve que certains estimateurs indirects (voir le chaptitre V) ont la bonne propriété de converger vers un estimateur direct quand la taille d'échantillon augmente. Les estimateurs indirects ayant cette propriété peuvent de ce fait être utilisés pour tous les domaines, qu'ils soient "petits" ou pas, ce qui relativise l'importance d'avoir une définition formelle plus ou moins exacte de ce qu'est un « petit domaine ».

En pratique, on doit cependant spécifier le niveau de désagrégation pertinent pour continuer à utiliser des estimateurs conventionnels directs, le niveau auquel passer à des estimateurs indirects, et même, si c'est approprié, être capable de produire des données statistiques à tout niveau de désagrégation, ce qui à la limite pourrait conduire jusqu'au niveau individuel. En fonction de ce qui a été écrit précédemment, il est conseillé d'utiliser des estimateurs directs au niveau auquel les CVs de ces estimateurs ne dépassent pas la limite fixée pour l'ensemble des domaines. Si cette limite est dépassée, il est préférable d'utiliser des estimateurs indirects pour tous les domaines de ce niveau.

Il faut indiquer ici qu'il n'est pas conseillé de produire des estimations pour tous les domaines, si le modèle ne peut être vérifié (pratiquement, aucun modèle n'est parfaitement vrai), car le biais introduit par des estimateurs indirects augmente quand la taille d'échantillon diminue. Bien que l'erreur quadratique moyenne des estimateurs indirects reste plus faible que celle des estimateurs directs, il n'est pas conseillé de s'abstraire totalement du biais relatif à ces méthodes. Il est donc intéressant de fixer une limite supérieure pour le biais relatif d'un estimateur indirect et choisir de ne pas publier des données pour les domaines où cette limite est dépassée. Cette limite sera établie en fonction des besoins des utilisateurs des données (par exemple 10% ou 5% du biais relatif absolu). De ce fait, il est recommandé :

- D'utiliser des estimateurs directs pour la population entière, et à des niveaux élevés d'agrégation, si les estimateurs directs ont un CV en-dessous de la limite fixée pour l'ensemble des domaines.
- Pour des niveaux plus désagrégés, d'utiliser des estimateurs indirects pour les domaines où le biais relatif absolu ne dépasse pas une valeur fixée a priori.

- Enfin, pour les domaines où les estimateurs indirects ont un biais relatif absolu supérieur à la valeur pré-fixée, il est conseillé de ne pas produire d'estimations, ou de modifier le plan de sondage de l'enquête de façon à disposer des tailles d'échantillon mínimum dans chaque domaine d'intérêt.

Le biais d'un estimateur ne peut pas être connu de manière certaine, car il dépend de la vraie valeur, inconnue, de l'indicateur considéré. Mais dans certains cas il peut être approché de manière théorique. Une autre option est de l'évaluer de manière empirique au travers de simulations. Celles-ci peuvent être réalisées en simulant des données de la manière la plus complète possible, par exemple à partir de données d'un recensement, ou en utilisant des données d'enquête servant à générer un recensement dans lequel on va tirer des échantillons. Ce type d'expérimentation a un intérêt très important, qui est qu'il permet de valider les méthodes d'estimation, dans les situations où les vraies valeurs sont connues. Dans les deux cas indiqués ci-dessus, il est possible de déterminer la taille d'échantillon minimum requise au niveau domaine qui conduit à ne pas dépasser une valeur supérieure du biais absolu relatif pour l'estimateur de l'indicateur considéré (voir section V.A). C'est pourquoi, pour produire des estimations, les estimateurs indirects ne devraient être utilisés que dans les domaines où la taille d'échantillon dépasse cette valeur minimum.

C. Méthodes destinées à résoudre les problèmes posés par la désagrégation des données

Comme indiqué précédemment, si l'objectif est d'éviter des augmentations de tailles d'échantillon en raison des conséquences sur les coûts, ou si la demande de données désagrégées à un niveau fin arrive après que l'enquête a été réalisée, une manière économique d'obtenir des estimations pour des domaines d'intérêt qui sont plus fiables que celles obtenues à partir des estimateurs directs consiste à utiliser des estimateurs indirects. Ces dernières méthodes n'utilisent pas seulement les données d'enquête relatives au domaine considéré, mais également des données d'autres domaines, qui ont une certaine ressemblance avec le domaine considéré. Cette ressemblance est généralement représentée par l'intermédiaire d'un modèle (avec un ensemble d'hypothèses). Les estimateurs indirects les plus simples sont basés sur des hypothèses irréalistes, et peuvent de ce fait avoir des biais très importants. Ils comprennent des estimateurs synthétiques, qui ne prennent pas en compte l'hétérogénéité existant habituellement entre domaines. Les estimateurs synthétiques populaires sont l'estimateur synthétique post-stratifié et l'estimateur synthétique de type régression (chapitre III). Comme autres estimateurs indirects classiques, on trouve les estimateurs composites, calculés comme la somme pondérée entre un estimateur direct et un indicateur synthétique, qui incluent l'estimateur dépendant de la taille d'échantillon et/ou les estimateurs composites optimaux. La pondération donnée à chaque estimateur ne dépend pas de la qualité de l'ajustement du modèle utilisé par l'estimateur synthétique. De plus, en pratique le poids de l'estimateur direct est souvent proche de 1, ce qui implique que peu d'information est « empruntée » en complément.

Des estimateurs indirects légèrement plus sophistiqués, qui prennent en compte l'existence de différences entre les domaines, sont ceux basés sur des modèles de régression. Il y a deux principaux groupes de modèles de régression utilisés pour l'estimation sur petits domaines : les modèles au niveau domaine, et les modèles au niveau individu, et il est également possible de proposer des modèles à des niveaux intermédiaires d'agrégation (par exemple, par groupes de sexe/âge à l'intérieur de zones géographiques). Les modèles au niveau domaine utilisent seulement des données agrégées pour les domaines (ou zones) sur lesquels on veut produire des estimations. Ce type de données est facile à obtenir, car l'agrégation évite d'être confronté à des problèmes de confidentialité. Les modèles dits de Fay-Herriot (FH), proposés par Fay et Herriot (1979), sont des modèles linéaires largement répandus pour l'estimation sur petits domaines. Ces modèles ont une structure à deux niveaux.

Au premier niveau, la relation entre les indicateurs d'intérêt pour les domaines et les variables auxiliaires au niveau domaine est supposée la même pour tous les domaines. Par exemple, la perte de revenu quand on passe de l'emploi au chômage est supposée, toutes choses égales par ailleurs, la même dans tous les domaines. Ceux-ci sont donc reliés par un modèle de régression linéaire. Au second niveau, il est supposé que, les vraies valeurs de la variable d'intérêt étant données, les estimateurs directs sont centrés sur ces vraies valeurs et ont des variances supposées connues. Ces variances varient entre les domaines, étant donné que les tailles des échantillons qui s'y trouvent sont différentes. Ces modèles sont populaires à juste titre, car les estimateurs résultants sont une moyenne pondérée (composite) des estimateurs directs et des estimateurs synthétiques par la régression. Quand le modèle synthétique ne s'ajuste pas bien aux données (si les variables auxiliaires considérées n'expliquent pas suffisamment bien l'hétérogénéité de l'indicateur entre les domaines), ou quand la taille de l'échantillon dans un domaine est grande, l'estimateur FH donne plus de poids à l'estimateur direct, qui est suffisamment précis. A l'inverse, quand le modèle synthétique s'ajuste bien ou quand la taille d'échantillon est petite (donc avec un estimateur direct imprécis), le poids accordé à l'estimateur synthétique de type régression augmente. Das ce cas, l'efficacité est améliorée, car l'estimateur synthétique a un coefficient de régression qui est commun à tous les domaines et qui peut donc être estimé à partir de l'ensemble des données des domaines. De plus, puisque les estimateurs directs sont approximativement sans biais sous le plan de sondage, quand on se trouve dans un domaine avec une grande taille d'échantillon les estimateurs obtenus à partir du modèle FH ont également un biais limité sour le modèle. Un des problèmes posés par ces modèles est de déterminer les valeurs des variances des estimateurs directs (ou variances hétéroscédastiques des termes d'erreur du modèle). Bien que, comme indiqué précédemment, on suppose que les variances sont connues, dans la pratique on les remplace par des estimations. Et étant donné le faible nombre de données dans certains domaines, les estimations de ces variances sont également de mauvaise qualité. Il existe des méthodes de lissage comme la méthode de la fonction généralisée de la variance (voir Fay et Herriot, 1979) ou l'estimation non paramétrique de ces variances (voir González-Manteiga et al. (2010)). Cette estimation vient ajouter un problème, qui est celui de l'incorporation de cette erreur dans l'erreur de l'estimateur final.

Dans les modèles au niveau individu, comme leur nom l'indique, le modèle est construit pour chaque individu de la population (modèle de superpopulation), et son ajustement demande des données individuelles sur la variable de réponse et les variables auxiliaires. Le premier modèle de ce type a été proposé par Battesse, Harter et Fuller (1988) et est connu comme modèle à erreur emboitée. C'est un modèle de régression linéaire qui, en plus des erreurs de modèle individuelles, intègre des effets aléatoires associés aux domaines, qui expliquent l'hétérogénéité entre domaines non prise en compte par les variables auxiliaires. Ces modèles sont largement utilisés quand les données nécessaires sont disponibles, parce qu'ils incorporent beaucoup plus d'information que les modèles au niveau domaine, et que les variances des erreurs du modèle ne nécessitent pas d'être connues.

L'introduction d'un modèle stochastique générant les valeurs de la variable d'intérêt dans les individus de la population fait des indicateurs d'intérêt des quantités aléatoires. C'est pourquoi il est courant, dans la littérature dédiée, d'utiliser le terme « prédire » plutôt qu' « estimer » la valeur d'un indicateur d'intérêt, et « prédicteur » plutôt qu' « estimateur ». Dans ce document, les deux termes seront utilisés comme des synonymes. Dans ce contexte, un prédicteur sans biais d'un indicateur est tel que son espérance sous le modèle est égale à l'espérance de cet indicateur. Quand on estime des indicateurs obtenus de façon linéaire à partir des valeurs individuelles d'une variable d'intérêt, comme des moyennes ou des totaux, les modèles de base utilisés au niveau domaine ou individu font partie de modèles mixtes linéaires incluant des effets aléatoires sur les domaines d'intérêt. Pour ces modèles, l'estimateur usuel indirect est le meilleur prédicteur linéaire sans biais (en anglais: best linear unbiased predictor - BLUP), qui consiste en la combinaison linéaire des valeurs observées pour la variable sur les individus de l'échantillon, et qui est sans biais sous le modèle et minimise l'erreur quadratique moyenne. Le BLUP dépend des paramètres inconnus du modèle, qui représentent le comportement commun

entre domaines. En remplaçant les paramètres inconnus par des estimations, on obtient un BLUP empirique (EBLUP). Ce peut être l'estimateur (prédicteur) usuel basé sur un modèle d'un indicateur linéaire sur un petit domaine.

Le BLUP ne nécessite aucune hypothèse de normalité dans le modèle. Cependant, pour estimer des indicateurs plus généraux que les indicateurs linéaires, le meilleur prédicteur est celui qui minimise l'erreur quadratique moyenne, sans qu'il soit nécessaire qu'il soit linéaire ou sans biais. Ce qui est équivalent à l'espérance sous le modèle de l'indicateur à estimer, conditionnellemet aux valeurs observées sur l'échantillon. Sous l'hypothèse normale, le meilleur prédicteur d'un indicateur linéaire est le BLUP. Quand il n'y a pas de normalité, ou quand l'indicateur à estimer n'est pas linéaire, il peut arriver que l'espérance définissant le meilleur prédicteur soit impossible à calculer de manière analytique. Dans ce cas, des approximations numériques du meilleur prédicteur sont utilisées. D'autres modèles largement utilisés, par exemple lorsque l'on estime des proportions ou des variables binaires, sont les modèles linéaires généralisés à effets aléatoires (voir le chapitre V).

Nous considérons maintenant une population divisée en domaines, ces domaines étant à leur tour divisés en sous-domaines, et souhaitons produire des estimations à un ou deux de ces niveaux. Par exemple, le Mexique est divisé en 31 Etats plus Mexico City et chaque Etat, à son tour, est divisé en un certain nombre de municipalités. Les modèles les plus appropriés pour ce type de contexte incluent des effets aléatoires aux différents niveaux (voir, par exemple, Stukel et Rao, 1999 pour l'estimation des indicateurs linéaires ou Marhuenda et al., 2018, pour l'estimation d'indicateurs généraux). D'une part, quand il y a plusieurs variables d'intérêt qui sont liées, des modèles multivariés peuvent être utilisés (voir Fay, 1987 ou Datta, Fay et Ghosh, 1991). Egalement, quand il existe une corrélation temporelle et/ou spatiale, on peut recourir à des modèles qui incluent des effets aléatoires suivant un processus de séries temprelles et/ou un processus spatial (voir, par exemple, Pfeffermann et Burk (1990) ou Rao et Yu (1992) pour les modèles temporels, Molina, Salvati et Pratesi (2008) pour un modèle spatial et Marhuenda, Molina et Morales (2013) pour un modèle spatio-temporel). D'autre part, les modèles bayésiens constituent une alternative aux modèles fréquentiels qui présente souvent des avantages en temps de calcul, fournissant des estimateurs pratiquement identiques à ceux obtenus avec le modèle fréquentiel correspondant, tant que la distribution a priori considérée est non informative (voir chaptitre IV). L'étude de cas par Rao et Molina (2015) donne un aperçu détaillé des techniques les plus largement utilisées pour l'estimation sur petits domaines, et constitue une revue complète de la plupart des travaux menés sur le sujet au moment de sa publication.

II. Indicateurs usuels de pauvreté et d'inégalité

Il existe beaucoup d'indicateurs de pauvreté et d'inégalité qui résument les différents aspects des conditions de vie d'une population. En effet, à partir des enquêtes statistiques officielles sur les conditions de vie réalisées dans de nombreux pays, les Instituts Nationaux Statistiques produisent, de manière régulière, une grande variété d'indicateurs destinés à donner des éléments de mesure de la pauvreté et des inégalités. Il est important de considérer la forme mathématique de chaque indicateur qu'on cherche à estimer, car quand on veut choisir des méthodes d'estimation sur petits domaines, toutes les techniques ne sont pas applicables à l'ensemble des indicateurs.

Dans ce chapitre, nous passons en revue beaucoup des indicateurs disponibles dans la littérature, ainsi que les indicateurs produits habituellement à partir des enquêtes officielles sur les conditions de vie. Même s'il n'est pas possible d'inclure l'ensemble de ces indicateurs, un certain nombre de ceux décrits dans ce chapitre seront utilisés pour illustrer les techniques d'estimation sur petits domaines les plus utilisées. Dans les chapitres suivants, les différentes méthodes possibles seront donc présentées avec des éléments sur les types d'indicateurs auxquels elles sont adaptées.

Neri, Ballini et Betti (2005) passent en revue les indicateurs de pauvreté et d'inégalité. L'indicateur de pauvreté le plus répandu est l'incidence de la pauvreté, ou taux de pauvreté, encore appelé taux de risque de pauvreté, qui est calculé comme la proportion d'individus qui ont un revenu (équivalent net) en-dessous du seuil de pauvreté. Un autre indicateur usuel est l'écart de pauvreté, qui mesure l'étendue de la pauvreté plutôt que la fréquence des individus en risque de pauvreté. Ces deux indicateurs font partie d'une famille plus large d'indicateurs définie par Foster, Greer et Thorbecke (1984), que nous appellerons la famille d'indicateurs FGT, ces indicateurs ayant l'avantage d'être additifs pour les individus. Les méthodes d'estimation sur petits domaines que nous décrirons dans les chapitres suivants seront illustrées en les appliquant à ces indicateurs, mais il est important de remarquer que certaines de ces méthodes sont applicables à beaucoup d'autres indicateurs ne faisant pas partie de la famille FGT. Dans chaque chapitre, on fera apparaître clairement à quels indicateurs chaque méthode est applicable.

On note U la population cible (par exemple les résidents d'un pays), de taille N, qui est divisée en D sous-populations, les domaines (ou zones) à estimer, de tailles $N_1, \dots, N_D$. On peut noter que les tailles (de population) des domaines sont en général grandes puisque, comme indiqué au chapitre I, le terme "petit domaine" se réfère à la taille de l'échantillon (plus exactement, à la valeur de l'erreur d'échantillonnage de l'estimateur direct) et non à la taille de la population.

E_{di} est la mesure du pouvoir d'achat (par exemple, le revenu, ou la dépense totale) de l'individu i dans le domaine d, $d = 1, \dots, D$. z est le seuil de pauvreté utilisé, en-dessous duquel un individu est considéré comme en risque de pauvreté. La famille d'indicateurs FGT pour le domaine d est définie comme :

$$F_{\alpha d} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left(\frac{z - E_{di}}{z} \right)^{\alpha} I(E_{di} < z), \quad d = 1, \dots, D, \alpha \geq 0, \tag{1}$$

où $I(E_{di} < z)$ est la fonction indicatrice, valant 1 si $E_{di} < z$ (individu i en risque de pauvreté) ou la valeur 0 sinon. Si l'on prend comme valeur $\alpha = 0$, on obtient le taux de pauvreté (ou incidence de la pauvreté). L'écart de pauvreté est l'indicateur obtenu avec $\alpha = 1$.

Un indicateur plus complexe qui utilise à la fois l'écart de pauvreté et l'incidence de la pauvreté, en plus du coefficient de Gini, est l'indicateur de Sen (Sen, 1976). Par ailleurs, parmi les indicateurs qui ne dépendent pas d'un seuil de pauvreté mais de la situation relative des individus au travers de leur classement, on peut mentionner le « Fuzzy monetary index » et le « Fuzzy supplementary index » (voir Betti et al. (2006)). Au-delà de la dimension monétaire, il est souvent intéressant de mesurer d'autres types de contraintes ou de privations qui ne sont pas strictement monétaires. Ces privations sont en général mesurées comme les proportions d'individus qui ont (ou n'ont pas) accès à certains services tels que les soins de santé, le logement, et l'éducation. Enfin, les indicateurs d'inégalité incluent l'indice de Gini, l'indice d'entropie généralisée, ou l'indice de Theil (voir par exemple Neri, Ballini et Betti (2005)).

Au Conseil Européen de décembre 2001, faisant partie de la Stratégie de Lisbonne de 2000 concernant la coordination des politiques sociales des Etats membres, un ensemble d'indicateurs de pauvreté et d'exclusion sociale, connu sous le nom d'indicateurs de Laeken, a été mis en place. Ces indicateurs comprennent le taux de risque de pauvreté F_{0d}, , le ratio des quintiles de revenu (le ratio entre les revenus des 20% les plus riches de la population et des 20% les plus pauvres), l'écart médian d'écart de pauvreté, et l'indice de Gini, entre autres.

Un exemple de mesure multidimensionnelle de la pauvreté est l'indicateur utilisé par le CONEVAL (Conseil National pour l'Evaluation de la Politique de Développement Social) au Mexique, connu comme l'indicateur multidimensionnel de pauvreté qui mesure la proportion d'individus qui sont concernés par au moins un des items parmi un ensemble donné de préjudices ou de privations, et dont le revenu est en-dessous d'un « seuil de bien-être ». Les chapitres suivants passent en revue les méthodes d'estimation sur petits domaines qui, bien que présentées au travers de leur application à la famille des indicateurs FGT, peuvent être utilisées de la même façon pour un ensemble plus large d'indicateurs.

III. Méthodes directes

Ce chapitre dresse un panorama des estimateurs directs de base pour la moyenne d'une variable sur un domaine, qui peut s'écrire

$$\bar{Y}_d = N_d^{-1} \sum_{i=1}^{N_d} Y_{di}, \qquad (2)$$

où Y_{di} est la valeur de la variable pour l'individu i situé dans le domaine (souvent une zone) d. Il faut remarquer que les indicateurs FGT spécifiés en (1) peuvent être écrits sous forme de moyennes, comme dans (2), en notant

$$F_{\alpha,di} = \left(\frac{z - E_{di}}{z}\right)^{\alpha} I(E_{di} < z),$$

ce qui permet d'écrire que $F_{\alpha d}$ est la moyenne des valeurs $Y_{di} = F_{\alpha,di}$ pour les individus du domaine d, ou en d'autres termes,

$$F_{\alpha d} = N_d^{-1} \sum_{i=1}^{N_d} F_{\alpha,di}. \qquad (3)$$

Comme indiqué précédemment, un estimateur d'un indicateur pour un domaine donné est qualifié de "direct" s'il est calculé en n'utilisant que les données de ces domaines, sans prendre en compte les données des autrres domaines. Ce type d'estimateur est l'estimateur utilisé par défaut par les Instituts Nationaux Statistiques, en raison de ses bonnes propriétés statistiques relativement à l'échantillonnage (en particulier son caractère sans biais) pour les domaines où la taille d'échantillon est suffisamment grande. Par exemple, les estimateurs directs ont été utilisés depuis longtemps pour produire des statistiques sur les conditions de vie au Chili au niveau national et régional, et pour un ensemble de communes ayant un échantillon representatif, à partir de l'enquête nationale chilienne de

« caractérisation socio-économique » (Encuesta de Caracterización Socioeconómica Nacional ou CASEN). A partir de la CASEN 2015, la méthodologie utilisée pour l'estimation dans les communes "non représentatives" a pris en compte des méthodes indirectes basées sur un modèle, plus précisément la méthode de Fay-Herriot décrite dans l'introduction (voir le papier sur la méthodologie d'estimation de la pauvreté au niveau commune, avec les données CASEN 2015 de l'Observatoire social chilien du Ministère du Développement Social, 2017).

Dans ce papier, s est l'échantillon de taille n tiré dans la population U, s_d le sous-échantillon du domaine d de taille n_d (qui peut être égale à zéro) et r_d l'ensemble des unités du même domaine situées en-dehors de l'échantillon, $d = 1, \dots, D$, avec $\sum_{d=1}^{D} n_d = n$. De plus, π_{di} est la probabilité d'inclusion de l'individu i dans l'échantillon du domaine d, $w_{di} = \pi_{di}^{-1}$ est le poids de sondage du même individu et $\pi_{d,ij}$ est la probabilité d'inclusion des individus i et j dans l'échantillon du domaine d. Nous allons maintenant présenter un panorama des estimateurs directs les plus connus.

A. Les estimateurs directs basiques

L'estimateur sans biais, sous le plan de sondage, de la moyenne d'une variable dans un domaine d, $\bar{Y}_d$, est connu sous le nom d'estimateur de Horvitz-Thompson (HT). Cet estimateur nécessite de connaitre la vraie valeur de la taille du domaine N_d et les poids d'échantillonnage $w_{di} = \pi_{di}^{-1}$ pour les individus échantillonnés dans le domaine d. Si l'on suppose qu'ils sont connus, l'estimateur HT de $\bar{Y}_d$ vaut

$$\hat{\bar{Y}}_d = N_d^{-1} \sum_{i \in s_d} w_{di}\, Y_{di}. \tag{4}$$

Notons que pour le total sur le domaine d, $Y_d = \sum_{i=1}^{N_d} Y_{di}$, l'estimateur HT est tout simplement $\hat{Y}_d = \sum_{i \in s_d} w_{di}\, Y_{di}$ et ne demande pas de connaître les tailles des domaines N_d.

Si $\pi_{di} > 0$ pour chaque $i = 1, \dots, N_d$, un estimateur sans biais de la variance de l'estimateur HT de $\bar{Y}_d$ peut s'écrire

$$\widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d) = N_d^{-2} \left\{ \sum_{i \in s_d} \frac{Y_{di}^2}{\pi_{di}^2}(1 - \pi_{di}) + 2 \sum_{i \in s_d} \sum_{\substack{j \in s_d \\ j > i}} \frac{Y_{di} Y_{dj}}{\pi_{di}\pi_{dj}} \left(\frac{\pi_{d,ij} - \pi_{di}\pi_{dj}}{\pi_{d,ij}} \right) \right\}. \tag{5}$$

Il arrive souvent que, quand on se situe à l'étape d'estimation, on ne dispose pas de toute l'information sur l'échantillonnage, à l'exception des poids d'échantillonnage w_{di}. Du fait que les probabilités d'inclusion double $\pi_{d,ij}$ ne sont pas disponibles, l'estimateur (5) ne peut pas être calculé. Cependant, pour des plans de sondage avec des probabilités d'inclusion double vérifiant $\pi_{d,ij} \approx \pi_{di}\pi_{dj}$, pour $j \neq i$, comme par exemple le sondage de Poisson, pour lequel l'égalité est vraie, le deuxième terme de (5) devient approximativement zéro. De plus, en utilisant $w_{di} = \pi_{di}^{-1}$, nous obtenons l'estimateur de variance suivantt, qui ne dépend pas des probabilités d'inclusion double

$$\widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d) = N_d^{-2} \sum_{i \in s_d} w_{di}\,(w_{di} - 1) Y_{di}^2. \tag{6}$$

Cet estimateur est fourni dans la fonction direct () du package R sae, qui sera utilisé dans l'exemple 1 pour illustrer cette procédure, quand les poids de sondage sont intégrés. Cette fonction part de l'hypothèse qu'aucune information sur le plan de sondage autre que les poids d'échantillonnage n'est disponible. Si l'on a plus d'information sur le plan de sondage, il est préférable d'utiliser des packages R plus appropriés tels que survey (Lumley 2017) ou sampling (Tillé and Matei 2016). De plus, il existe des approximations alternatives de la variance prenant en compte le plan de sondage et l'information

disponible, par exemple la méthode des « ultimate clusters » ou la méthode BRR (*Balanced Repeated Replications)* avec la correction de Fay (U.S. Bureau of Labor Statistics et U.S. Census Bureau 2006).

L'estimateur HT pondère les observations individuelles Y_{di} en utilisant les poids de sondage, ou inverses des probabilités d'inclusion, $w_{di} = \pi_{di}^{-1}$. Ceci évite les situations où la probabilité de sélectionner un individu est liée à la valeur de la variable d'intérêt (plan d'échantillonnage informatif). En effet, si certains individus (par exemple ceux avec un bas revenu) ont une probabilité plus forte d'être sélectionnés dans l'échantillon, ce type d'invidus apparaîtra plus souvent dans l'échantillon final, alors que les individus ayant moins de chances d'être sélectionnés (par exemple ceux à haut revenu) seront rarement dans l'échantillon. Ce qui veut dire que si l'on voulait passer par une estimation donnant le même poids à toutes les observations de l'échantillon, comme la moyenne simple calculée sur l'échantillon, on aurait un biais (en l'occurrence, le revenu moyen serait sous-estimé). Pour cette raison, un poids moindre doit être donné aux observations qui ont le plus de chances d'être sélectionnées, et un poids plus élevé à celles qui ont moins de chances d'être dans l'échantillon.

Bien que cet estimateur soit exactement sans biais relativement au plan de sondage, sa variance peut être très grande quand la taille de l'échantillon dans le domaine n_d est petite. Un estimateur légèrement biaisé pour les petites valeurs de n_d , mais avec une variance relativement plus faible, et qui ne nécessite pas de connaître la taille du domaine N_d pour estimer la moyenne $\bar{Y}_d$, est l'estimateur de Hájek. Cet estimateur est égal à la moyenne pondérée des observations du domaine, qui utilise les poids de sondage comme pondérations,

$$\hat{\bar{Y}}_d^{HA} = \widehat{N}_d^{-1} \sum_{i \in s_d} w_{di}\, Y_{di}, \qquad \text{avec } \widehat{N}_d = \sum_{i \in s_d} w_{di}.$$

Pour le total $Y_d = \sum_{i=1}^{N_d} Y_{di}$, l'estimateur de Hájek vaut $\hat{Y}_d^{HA} = N_d \hat{\bar{Y}}_d^{HA}$, ce qui nécessite de connaître la taille de la population N_d.

Sous le plan de sondage, un estimateur de la variance de l'estimateur de Hájek, $\hat{\bar{Y}}_d^{HA}$, est obtenu grâce à la méthode de linéarisation de Taylor. L'estimateur qui en résulte est obtenu en remplaçant Y_{di} par $\tilde{e}_{di} = Y_{di} - \hat{\bar{Y}}_d^{HA}$ dans l'estimateur de variance de l'estimateur HT du total $\hat{Y}_d$ et en divisant par $\widehat{N}_d$; à savoir

$$\widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d) = \widehat{N}_d^{-2} \left\{ \sum_{i \in s_d} \frac{(Y_{di} - \hat{\bar{Y}}_d^{HA})^2}{\pi_{di}^2} (1 - \pi_{di}) \right.$$

$$\left. + 2 \sum_{i \in s_d} \sum_{\substack{j \in s_d \\ j > i}} \frac{(Y_{di} - \hat{\bar{Y}}_d^{HA})(Y_{di} - \hat{\bar{Y}}_d^{HA})}{\pi_{di}\pi_{dj}} \left(\frac{\pi_{d,ij} - \pi_{di}\pi_{dj}}{\pi_{d,ij}} \right) \right\}, \qquad (7)$$

en supposant que $\pi_{di} > 0$, pour tout i. Pour les plans de sondage où $\pi_{d,ij} \approx \pi_{di}\pi_{dj}$, for $j \neq i$, par exemple l'échantillonnage de Poisson, cette variance estimée se réduit à

$$\widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d) = \widehat{N}_d^{-2} \sum_{i \in s_d} w_{di}\, (w_{di} - 1)(Y_{di} - \hat{\bar{Y}}_d^{HA})^2.$$

Comme indiqué précédemment, les indicateurs FGT ont l'avantage de pouvoir être écrits comme une moyenne des individus sur le domaine (voir (3)). De ce fait, l'estimateur de Horvitz-Thompson de $F_{\alpha d}$ vaut alors

$$\hat{F}_{\alpha d} = N_d^{-1} \sum_{i \in s_d} w_{di}\, F_{\alpha,di}.$$

De manière alternative, l'estimateur de Hájek de $F_{\alpha d}$ peut être écrit comme

$$\hat{F}_{\alpha d}^{HA} = \hat{N}_d^{-1} \sum_{i \in s_d} w_{di} F_{\alpha,di}.$$

Notons que, en agrégeant les estimateurs directs HT des totaux Y_d sur les domaines d'un ensemble plus large (par exemple une région, ou la population entière), nous obtenons l'estimateur HT de la population totale $\hat{Y} = \sum_{d=1}^{D} \sum_{i \in s_d} w_{di} Y_{di}$, ce qui veut dire que

$$\sum_{d=1}^{D} \hat{Y}_d = \hat{Y}.$$

Etant donné que l'estimateur HT est efficace à un niveau élevé d'agrégation, tel que la population entière, cette propriété, connue en anglais sous l'appellation "benchmarking property" (propriété d'additivité, ou de réconciliation des données), est recommandée pour l'estimation sur les domaines. Mais d'autres estimateurs, en particulier les estimateurs indirects présentés dans les chapitres suivants, ne sont pas parfaitement additifs pour retrouver l'estimateur direct considéré pour la population totale (et qui peut être différent de l'estimateur HT). Des ajustements peuvent alors être opérés pour forcer cette aditivité. Soit $\hat{Y}_d^{EST}$ un estimateur qui ne vérifie pas cette propriété. Si nous voulons que les estimateurs par domaine s'agrègent pour retrouver l'estimateur HT au niveau national $\hat{Y}$, un ajustement courant est celui qui passe par un ratio, et qui s'écrit

$$\hat{Y}_d^{AEST} = \hat{Y}_d^{EST} \frac{\hat{Y}}{\sum_{d=1}^{D} \hat{Y}_d^{EST}}, \quad d = 1, \dots, D.$$

Il existe une littérature abondante sur d'autres méthodes d'ajustement, telles que les ajustements par la différence, et sur des méthodes destinées à contraindre les estimateurs à respecter cette propréiét à différents niveaux, et qui ne sont pas présentées dans ce document faute de place. Pour plus d'informations, on pourra se référer à Ghosh et Steorts (2013) et à la bibliographie qui y est donnée.

Ci-dessous, nous résumons les types d'indicateurs auxquels ces estimateurs sont applicables, les données qui sont nécessaires pour les produire, comme les données concernant la variable d'intérêt obtenues via l'enquête, et les avantages et inconvénients tels que perçus de façon pratique.

Indicateurs cibles : paramètres additifs, en ce sens qu'ils sont la somme de certaines variables pour chaque individu du domaine. Ces variables peuvent être des fonctions de variables d'intérêt des individus (par exemple, $F_{\alpha,di}$ est une fonction de la variable utilisée pour mesurer le pouvoir d'achat de l'individu, E_{di}).

Données requises:

- Poids de sondage w_{di} pour les individus échantillonnés dans le domaine d.
- Pour l'estimateur HT de la moyenne ou pour l'estimateur de Hájek du total, taille de la population dans chaque domaine, N_d.

Avantages:

- L'estimateur HT est exactement non biaisé et l'estimateur de Hájek est approximativement sans biais sous le plan de sondage. Tous les deux sont convergents par rapport au plan de sondage quand la taille de l'échantillon dans le domaine n_d augmente. C'est pouquoi ils ont de bonnes performances pour des domaines avec une taille d'échantillon suffisante dans le cas de plans de sondage où les probabilités de tirage

sont inégales, y compris les plans informatifs, tant qu'ils sont calculés avec les véritables probabilités d'inclusion des individus dans le domaine échantillonné.

- Ils ne nécessitent pas de modèle ou d'hypothèses sur les variables en question Y_{di} ce qui signifie qu'ils sont complètement non-paramétriques.
- Ils respectent la propriété d'additivité ("benchmarking property"): si l'on additionne les totaux estimés pour tous les domaines d'un ensemble (par exemple une région), on obtient exactement le total estimé pour cet ensemble par la même méthode.

Inconvénients:

- Ils sont très inefficaces (à savoir qu'is ont une erreur d'échantillonnage élevée) pour les petits domaines en raison de la faible taille d'échantillon.
- Ils ne peuvent pas être calculés pour les domaines non échantillonnés, ceux pour lesquels la taille d'échantillon n_d est égale à zéro.

Exemple 4.1. Estimateurs HT directs de l'incidence de la pauvreté, avec R. Nous allons montrer comment calculer des estimateurs HT directs pour l'incidence de la pauvreté, à partir de données simulées sur les conditions de vie dans les provinces espagnoles, contenues dans le fichier R appelé incomedata dans le package R sae. Ce fichier comprend, pour $n = 17119$ individus fictifs vivant dans $D = 52$ provinces espagnoles, le nom de la province où ils vivent (provlab), le code de la province (prov), le code de la communauté autonome (ac), le groupe d'âge de 1 à 5 (age), la nationalité (nat, 1=espagnole, 2=autre), le niveau d'éducation (educ, de 0=moins de 16 ans à 2=troisième niveau), le statut vis-à-vis de l'emploi (labor, où 0=moins de 16 ans, 1=employé, 2=chômage and 3=inactif), s'ils sont dans chaque groupe d'âge, pour les groupes 2 à 5 (age2 à age5), s'ils ont les niveaux d'éducation 1 à 3 (educ1 à educ3), s'ils ont la nationalité espagnole, s'ils sont employés, au chômage ou inactif, leur revenu équivalent net (income) et le poids de sondage (weight). Nous calculons les estimateurs HT directs de l'incidence de la pauvreté pour les $D = 52$ provinces espagnoles.

Après avoir installé la librairie R sae, we chargeons la base de données incomedata, qui contient les données échantillonnées, et la base de données sizeprov, qui contient les tailles de population pour les provinces, $\mathrm{N_d}$:

```
library(sae)
data(incomedata)
attach(incomedata)
data(sizeprov)
```

Ensuite, nous utilisons la fonction direct() pour obtenir les estimateurs HT directs. Tout d'abord, nous calculons la taille totale de l'échantillon, le nombre de provinces et leurs tailles d'échantillons, et extrayons les tailles de population du fichier sizeprov :

```
n<-dim(incomedata)[1]        # Taille totale d'échantillon
D<-length(unique(prov))      # Nombre de provinces (domaines, ici des zones)
nd<-as.vector(table(prov))   # Tailles d'échantillons des provinces
Nd<-sizeprov$Nd              # Tailles des populations des provinces
```

Nous fixons le seuil de pauvreté, calculé comme 0.6*valeur médiane du revenu à partir des données de l'année précédente, et créons la variable poor, qui indique si l'on a un revenu en-dessous du seuil de pauvreté:

```
z<-6557.143
poor<-numeric(n)
poor[income<z]<-1
```

Enfin, nous calculons les estimateurs HT directs de l'incidence de la pauvreté dans les provinces (moyennes de la variable poor dans les provinces), en utilisant la fonction *direct()* qui prend en compte les poids d'échantillonnage donnés par la variable *weight*:

```
povinc.dir.res<-direct(y=poor,dom=prov,sweight=weight,domsize=sizeprov[,-1])
print(povinc.dir.res,row.names=F)
```

Le résultat fourni par cette fonction est:

```
Domain SampSize     Direct         SD       CV
     1       96 0.25503732 0.04846645 19.003670
     2      173 0.14059242 0.03042195 21.638397
     3      539 0.20785096 0.02178689 10.481979
     4      198 0.26763976 0.04090335 15.282986
     5       58 0.05512200 0.02555426 46.359465
     6      494 0.21553890 0.02357906 10.939585
     7      634 0.09999792 0.01536517 15.365488
     8     1420 0.29812535 0.01618508  5.428952
     9      168 0.21413150 0.04473542 20.891562
    10      282 0.27031324 0.03125819 11.563692
    11      398 0.14887351 0.02189022 14.703904
    12      118 0.17598199 0.03584882 20.370731
    13      250 0.20921534 0.03279230 15.673948
    14      224 0.29975708 0.03934080 13.124228
    15      495 0.25347550 0.02467716  9.735520
    16       92 0.26334059 0.05913385 22.455274
    17      142 0.18337421 0.03710194 20.232911
    18      208 0.31727340 0.04043964 12.745990
    19       89 0.17908182 0.04234025 23.642966
    20      285 0.23690549 0.03194779 13.485457
    21      122 0.12583449 0.03202547 25.450474
    22      115 0.24107606 0.04856351 20.144476
    23      232 0.31294198 0.04122671 13.173916
    24      218 0.18801572 0.03002634 15.970122
    25      130 0.15559590 0.03872448 24.887854
    26      510 0.25811811 0.02459196  9.527405
    27      173 0.37718722 0.05696330 15.102129
    28      944 0.18218209 0.01639018  8.996593
    29      379 0.22918462 0.02735631 11.936364
    30      885 0.17703167 0.01648910  9.314210
    31      564 0.16190765 0.01842017 11.376958
    32      129 0.22799612 0.04199465 18.419018
    33      803 0.26064010 0.02093779  8.033220
    34       72 0.30166074 0.07179782 23.800849
    35      472 0.16651843 0.02307258 13.855869
```

```
36      448 0.18549072 0.02418887 13.040474
37      164 0.16104513 0.02998243 18.617410
38      381 0.18429619 0.02054550 11.148085
39      434 0.34244429 0.03248937  9.487491
40       58 0.22262002 0.05639965 25.334492
41      482 0.20503036 0.02122527 10.352256
42       20 0.02541207 0.02540651 99.978151
43      134 0.32035438 0.04934077 15.401934
44       72 0.27364239 0.06723440 24.570172
45      275 0.12553377 0.02131991 16.983409
46      714 0.21360678 0.02070508  9.693081
47      299 0.19292332 0.03211484 16.646429
48      524 0.21694466 0.02215645 10.212948
49      104 0.30027442 0.06025302 20.065986
50      564 0.10034577 0.01569138 15.637311
51      235 0.19724796 0.03341193 16.939048
52      180 0.19109119 0.03441016 18.007191
```

Pour finir, nous conservons les valeurs estimées dans un vecteur, et comptons combien de provinces ont un CV supérieur à 20%:

```
povinc.dir<-povinc.dir.res$Direct
povinc.dir.cv<-povinc.dir.res$CV
sum(povinc.dir.cv>20)
```

Il y a 15 provinces pour lesquelles les estimateurs HT directs de l'incidence de la pauvreté ont un CV supérieur à 20%. Ces 15 provinces sont considérées comme des petits domaines pour cet indicateur. Mais, comme on va le voir, des estimateurs plus efficaces peuvent également être trouvés pour les autres provinces.

B. GREG et estimateurs par calage

Un estimateur plus sophistiqué que les estimateurs directs de base présentés dans la partie précédente, car utilisant de l'information auxiliaire, est l'estimateur par la régression généralisée (GREG). Cet estimateur nécessite de connaitre le total $\boldsymbol{X}_d = \sum_{i=1}^{N_d} \boldsymbol{x}_{di}$, ou la moyenne $\bar{\boldsymbol{X}}_d = N_d^{-1} \sum_{i=1}^{N_d} \boldsymbol{x}_{di}$, pour le domaine d, d'un vecteur $\boldsymbol{x}_{di}$ de valeurs de p variables auxiliaires liées à Y_{di}, pour les individus i siués à l'intérieur du domaine d. Si $\hat{\bar{\boldsymbol{X}}}_d = N_d^{-1} \sum_{i \in s_d} w_{di}\, \boldsymbol{x}_{di}$ est l'estimateur HT de $\bar{\boldsymbol{X}}_d$, l'estimateur GREG de $\bar{Y}_d$ s'écrit

$$\hat{\bar{Y}}_d^{GREG} = \hat{\bar{Y}}_d + \left(\bar{\boldsymbol{X}}_d - \hat{\bar{\boldsymbol{X}}}_d\right)' \hat{\boldsymbol{B}}_d. \tag{8}$$

Ici, $\hat{\boldsymbol{B}}_d = \left(\sum_{i \in s_d} w_{di}\, \boldsymbol{x}_{di} \boldsymbol{x}_{di}'/c_{di}\right)^{-1} \sum_{i \in s_d} w_{di}\, \boldsymbol{x}_{di} Y_{di}/c_{di}$ est l'estimateur des moindres carrés pondérés (à partir des poids de sondage) du vecteur des coefficients de la régression linéaire supposée pour les unités du domaine d,

$$Y_{di} = \boldsymbol{x}_{di}' \boldsymbol{\beta}_d + \epsilon_{di}, \quad i = 1, \ldots, N_d, \tag{9}$$

où les erreurs du modèle ϵ_{di} sont indépendantes, avec une espérance nulle et une variance $\sigma^2 c_{di}$, les $c_{di} > 0$ étant des constantes représentant une possible hétéroscédasticité, $i = 1, \dots, N_d$. Les constantes c_{di} sont déterminées en étudiant les résidus du modèle linéaire sans hétéroscédasticité, c'est-à-dire avec $c_{di} = 1$, $i = 1, \dots, N_d$. Par exemple, en observant le nuage de points des résidus avec chacune des variables auxiliaires, on peut observer graphiquement si la variance des résidus augmente avec une d'entre elles. Dans ce cas, on prendra comme constantes c_{di}, les valeurs de cette variable pour les unités du domaine, ou, de manière plus générale, une fonction $c_{di} = f(x_{di}) > 0$, des valeurs de cette variable auxiliaire.

L'estimateur GREG de la moyenne sur le domaine d, $\bar{Y}_d$, est approximativement sans biais sous le plan de sondage que le modèle (9) soit valide ou pas, car le biais de l'estimateur du vecteur de coefficients de régression $\widehat{\boldsymbol{B}}_d$, en tant qu'estimateur de sa valeur sur la population, $\boldsymbol{B}_d = (\sum_{i=1}^{N_d} \boldsymbol{x}_{di}\, \boldsymbol{x}_{di}'/c_{di})^{-1} \sum_{i=1}^{N_d} \boldsymbol{x}_{di}\, Y_{di}/c_{di}$, est faible. C'est pourquoi le modèle (9) est souvent appelé « modèle de travail » et les estimateurs comme (8) qui sont sans biais, que le modèle soit valide ou non, sont appelés estimateurs assistés par un modèle. D'autre part, le GREG est également sans biais sous le modèle de régression (9), conditionnellement à l'échantillon s. Bien que l'estimateur GREG tende à améliorer l'efficacité de l'estimateur direct $\hat{\bar{Y}}_d$ si les variables auxiliaires sont liées par une relation linéaire à la variable dépendante Y_{di}, cet estimateur utilise uniquement des données du domaine d et, de ce fait, sa variance peut quand même être importante pour des domaines avec une petite taille d'échantillon n_d.

Notons que, si l'on veut utiliser l'estimateur GREG pour l'indicateur FGT d'ordre α, qui est égal à la moyenne des $F_{\alpha,di}$ dans le domaine, $F_{\alpha d} = N_d^{-1} \sum_{i=1}^{N_d} F_{\alpha,di}$, le gain en efficacité par rapport à l'estimateur direct dépendra de la qualité de l'ajustement du modèle de régression suivant :

$$F_{\alpha,di} = \boldsymbol{x}_{di}'\boldsymbol{\beta}_d + \epsilon_{di}, \quad i = 1, \dots, N_d.$$

Mais dans le cas des indicateurs FGT, les variables $F_{\alpha,di}$ sont une fonction complexe de la variable d'intérêt (la mesure du pouvoir d'achat E_{di}) exprimée comme $F_{\alpha,di} = \{(z - E_{di})/z\}^{\alpha} I(E_{di} < z)$, $\alpha \geq 0$. Il n'est pas facile de trouver des variables auxiliaires $\boldsymbol{x}_{di}$ qui sont liées de façon linéaire aux $F_{\alpha,di}$. C'est pourquoi ce modèle est difficile à être vérifié dans la pratique, ce qui fait que, pour les indicateurs FGT, les estimateurs GREG sont moins usités que pour des estimations de moyennes ou de totaux des variables d'intérêt (par exemple le revenu E_{di}).

Les estimateurs par calage sont largement utilisés par les Instituts Nationaux de Statistique pour estimer des moyennes ou des totaux au niveau national et pour les régions ayant une taille d'échantillon suffisante. Si l'on procède à un calage au niveau d'un domaine, l'estimateur résultat sera très proche de l'estimateur GREG. La méthode de calage a été proposée par Deville et Särndal (1992) pour estimer le total d'une variable d'intérêt en utilisant de l'information auxiliaire sur p variables. Si l'on suppose que l'on connait les totaux des variables auxiliaires sur un domaine $\boldsymbol{X}_d$, et si l'on suppose également que les variables auxiliaires $\boldsymbol{x}_{di}$ sont reliées de façon linéaire à Y_{di}, la méthode de calage consiste à trouver de nouveaux poids h_{di}, aussi proches que possible des poids de sondage d'origine w_{di}, au regard d'un mesure de distance $G_{di}(h_{di}, w_{di})$, de façon que le total $\boldsymbol{X}_d$ des variables auxiliaires soir estimé de façon parfaite avec ces nouveaux poids; donc sans erreur. Si la variable d'intérêt est liée de façon linéaire aux variables auxiliaires, et si les totaux de ces variables auxiliaries sont connus de manière exacte, les totaux de la variable d'intérêt seront estimés avec une petite erreur. En termes formels, quand on veut estimer la moyenne $\bar{Y}_d$, on cherche de nouveaux poids pour les unités échantillonnées, h_{di}, $i \in s_d$, qui sont la solution du problème suivant

$$\min_{\{h_{di}; i \in s_d\}} \quad \sum_{i \in s_d} G_{di}\,(h_{di}, w_{di})$$
$$\text{avec} \quad \sum_{i \in s_d} h_{di}\,\boldsymbol{x}_{di} = \boldsymbol{X}_d,$$

où $G_{di}(\cdot,\cdot)$ est une pseudo-distance. Si on utilise la distance pseudo chi-deux, qui s'écrit $G_{di}(h_{di}, w_{di}) = c_{di}(h_{di} - w_{di})^2/w_{di}$, et qui est probablement la distance la plus souvent adoptée, et si l'on résoud le problème grâce à la méthode du multiplicateur de Lagrange, les poids résultants sont

$$h_{di} = w_{di}\left\{1 + \boldsymbol{x}_{di}{}'\left(\sum_{i \in s_d} w_{di}\,\boldsymbol{x}_{di}\boldsymbol{x}_{di}{}'/c_{di}\right)^{-1}\left(\boldsymbol{X}_d - \sum_{i \in s_d} w_{di}\,\boldsymbol{x}_{di}/c_{di}\right)\right\}, i \in s_d. \qquad (10)$$

On peut noter que les poids calibrés h_{di} résultent d'un ajustement aux poids d'origine, $h_{di} = w_{di}g_{di}$, où le facteur d'ajustement g_{di} est le teme entre crochets dans (10). L'estimateur par calage de $\bar{Y}_d$ a donc la même forme que l'estimateur HT, mais avec des poids calibrés au lieu des poids d'origine:

$$\hat{\bar{Y}}_d^{CAL} = N_d^{-1} \sum_{i \in s_d} h_{di}\,Y_{di}.$$

Il est facile de montrer que, en substituant la formule obtenue pour les poids dans (10) dans l'estimateur par calage $\hat{\bar{Y}}_d^{CAL}$, on obtient exactement l'estimateur GREG de $\bar{Y}_d$ donné en (8). Deville et Särndal (1992) proposent des estimateurs par calage basés sur d'autres distances $G_{di}(\cdot,\cdot)$ que la distance du chi-deux. Mais ils montrent également que les estimateurs résultants, sous certaines conditions de régularité pour la distance $G_{di}(\cdot,\cdot)$, sont asymptotiquement équivalents au GREG et ont ainsi la même variance asymptotique. Comme pour l'estimateur GREG, pour une petite taille d'échantillon n_d, la variance des estimateurs par calage peut être grande.

Un estimateur consistant (quand la taille n_d augmente) pour la variance de l'estimateur $\hat{\bar{Y}}_d^{GREG}$ est obtenu en utilisant la méthode de linéarisation de Taylor. L'estimateur qui en résulte consiste à remplacer Y_{di} par $\tilde{e}_{di} = Y_{di} - \boldsymbol{x}_{di}{}'\hat{\boldsymbol{B}}_d$ dans la variance estimée de l'estimateur HT donnée en (5),

$$\widehat{\text{var}}_\pi(\hat{\bar{Y}}_d^{GREG}) = N_d^{-2}\left\{\sum_{i \in s_d} \frac{\tilde{e}_{di}^2}{\pi_{di}^2}(1 - \pi_{di}) + 2\sum_{i \in s_d}\sum_{\substack{j \in s_d \\ j > i}} \frac{\tilde{e}_{di}\tilde{e}_{di}}{\pi_{di}\pi_{di}}\left(\frac{\pi_{d,ij} - \pi_{di}\pi_{di}}{\pi_{d,ij}}\right)\right\}.$$

Pour les plans de sondage pour lesquels $\pi_{d,ij} \approx \pi_{di}\pi_{di}$ est vérifié, pour $j \neq i$, par exemple le plan de Poisson, cette variance estimée, écrite comme une fonction de $w_{di} = \pi_{di}^{-1}$, se réduit à

$$\widehat{\text{var}}_\pi(\hat{\bar{Y}}_d^{GREG}) = N_d^{-2} \sum_{i \in s_d} w_{di}\,(w_{di} - 1)\tilde{e}_{di}^2.$$

Des simulations ont montré que cet estimateur peut sous-estimer la variance du GREG. Mais l'estimateur obtenu en remplaçant Y_{di} par $g_{di}\tilde{e}_{di}$, où g_{di} est le facteur d'ajustment des poids w_{di}, dans la variance de l'estimateur HT, exprimée comme

$$\widehat{\text{var}}_\pi(\hat{\bar{Y}}_d^{GREG}) = N_d^{-2}\left\{\sum_{i \in s_d} \frac{g_{di}^2\tilde{e}_{di}^2}{\pi_{di}^2}(1 - \pi_{di}) + 2\sum_{i \in s_d}\sum_{\substack{j \in s_d \\ j > i}} \frac{g_{di}\tilde{e}_{di}g_{dj}\tilde{e}_{dj}}{\pi_{di}\pi_{di}}\left(\frac{\pi_{d,ij} - \pi_{di}\pi_{di}}{\pi_{d,ij}}\right)\right\}.$$

Réduit la sous-estimation et reste consistant quand n_d augmente (voir Fuller (1975) ou Estevao, Hidiroglou et Särndal (1995)). De plus, un tel estimateur alternatif de la variance est approximativement sans biais pour la variance du GREG $\hat{\bar{Y}}_d^{GREG}$ sous le modèle (9) conditionnellement à l'échantillon s, pour différents plans de sondage.

Remarquons à nouveau que ces estimateurs fonctionnent bien pour estimer des totaux ou des moyennes de variables d'intérêt, mais ne fonctionnent pas bien pour d'autres types de paramètres. Par exemple, pour l'indicateur FGT d'ordre α dans le domaine d, $F_{\alpha d} = N_d^{-1} \sum_{i=1}^{N_d} F_{\alpha,di}$, le GREG ou l'estimateur par calage seraient plus efficaces que l'estimateur direct si les variables auxiliaires $\boldsymbol{x}_{di}$ étaient liées de façon linéaire à $F_{\alpha,di}$,ce qui n'est pas le cas en pratique.

Les principales caractéristiques de ces estimateurs sont résumées ci-dessous:

Indicateurs cibles: moyennes/totaux de variables d'intérêt.

Données requises:

- Poids de sondage w_{di} pour les individus échantillonnés dans le domaine d.
- Pour l'estimateur de la moyenne, taille de la population dans le domaine, N_d.
- Observations sur l'échantillon des p variables auxiliaires liées à la variable d'intérêt, obtenues à partir de la même enquête que celle pour laquelle on dispose des variables d'intérêt.
- Totaux sur la population $\boldsymbol{X}_d$ ou moyennes $\bar{\boldsymbol{X}}_d$ des p variables auxiliaires dans le domaine.

Avantages:

- Les estimateurs sont approximativement sans biais (et consistants quand n_d augmente) sous le plan de sondage, que le modèle soit valide ou pas. C'est pourquoi ils ont de bonnes performances pour des domaines de taille d'échantillon suffisante et pour des plans de sondage avec probabilités inégales, y compris pour les plans de sondage informatifs.
- Ils ne demandent pas au modèle considéré d'être valide pour les variables d'intérêt Y_{di}; ils sont non paramétriques.

Inconvénients:

- Bien qu'ils puissent améliorer les estimateurs directs si le modèle de régression est bien vérifié, ils peuvent rester inefficaces pour des petits domaines en raison de leur petite taille d'échantillon.
- Ils ne peuvent pas être calculés pour des domaines (ou zones) où la taille d'échantillon n_d est égale à zéro.

Example 4.2. Estimateurs GREG de l'incidence de la pauvreté, avec R. Si l'on reprend l'exemple 4.1, nous montrons maintenant comment les estimateurs GREG de l'incidence de la pauvreté dans les provinces peuvent être calculés avec les mêmes données, mais en prenant en compte des variables auxiliaires; la variable constante égale à 1, le groupe d'âge, le niveau d'éducation, et le statut vis-à-vis de l'emploi.

Nous chargeons tout d'abord les fichiers comprenant les données nécessaires: les totaux des individus dans chaque province pour chaque groupe d'âge, pour chaque niveau d'éducation et pour chaque statut vis-à-vis de l'emploi:

```
data(sizeprovage)
data(sizeprovedu)
data(sizeprovlab)
```

Nous construisons la matrice avec les vecteurs de proportions des individus dans chaque catégorie et chaque province. Ce qui constitue le vecteur des moyennes sur la population $\bar{\boldsymbol{X}}_d$:

```
Nd<-sizeprov[,3]
Ndage<-as.matrix(sizeprovage[,-c(1,2)])
Ndedu<-as.matrix(sizeprovedu[,-c(1,2)])
Ndlab<-as.matrix(sizeprovlab[,-c(1,2)])

Pdage<-Ndage/Nd
Pdedu<-Ndedu/Nd
Pdlab<-Ndlab/Nd

X<-cbind(const=rep(1,D),Pdage[,3:5],Pdedu[,c(2,4)],Pdlab[,2])
```

Nous créons ensuite la matrice destinée à être utilisée pour la régression linéaire, avec les valeurs des variables auxiliaires pour les individus de l'échantillon :

```
Xtot<-model.matrix(poor~age3+age4+age5+educ1+educ3+labor1)
```

Enfin, nous calculons les estimateurs GREG pour l'incidence de la pauvreté (valeurs moyennes de la variable "pauvre") dans chaque province :

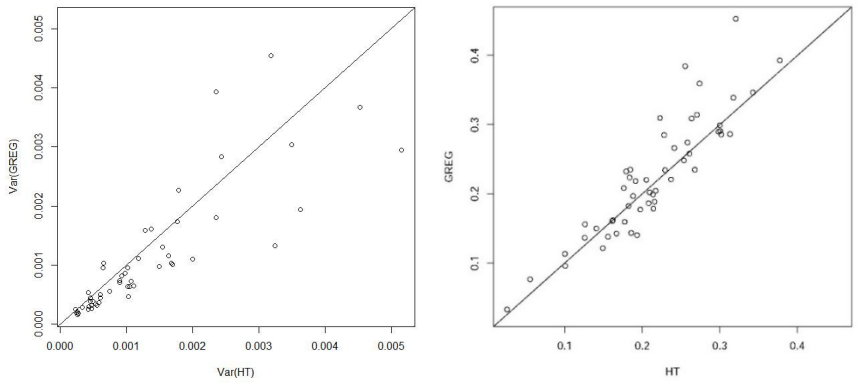
```
provl<-unique(prov)              # Indice de chaque province
p<-dim(Xtot)[2]                  # Nombre de variables auxiliaires
betad<-matrix(0,nr=D,nc=p)       # Matrice avec les coefficients de régression
                                 # pour chaque province (en ligne)
Xd.est<-matrix(0,nr=D,nc=p)      # Matrice des estimateurs directs des moyennes
                                 # des variables auxiliiaies pour chaque province
povinc.greg<-numeric(D)          # Vecteur des estimateurs GREG dans la province
povinc.greg.var<-numeric(D)      # Vecteur avec les variances estimées
                                 # pour les estimateurs GREG
for (d in 1:D){
  Xd<-Xtot[prov==provl[d],]      # Valeurs des variables auxiliaires
                                 # pour les individus de la province
  wd<-weight[prov==provl[d]]     # Poids de sondage des individus
                                 # dans la province
  yd<-poor[prov==provl[d]]       # Valeurs de la variable d'intérêt
                                 # pour les individus de la province

  # ajustement de la régression pour la province, avec les poids de sondage
  betad[d,]<-coef(summary(lm(yd~-1+Xd, weights=wd)))[,1]

  # Estimateurs directs des moyennes des variables auxiliaires dans la province
  Xd.est[d,]<-colSums(diag(wd)%*%Xd)/Nd[d]

  # Estimateur GREG de l'incidence de la pauvreté dans la province
  povinc.greg[d]<-povinc.dir [d]+sum((X[d,]-Xd.est[d,])*betad[d,])

  # Variance estimée pour l'estimateur GREG
  # de l'incidence de la pauvreté
  gd<-matrix(1/Nd[d]+
  +(X[d,]-Xd.est[d,])%*%solve(t(Xd)%*%diag(wd)%*%Xd)%*%t(Xd),nr=nd[d])
```

```
  ed<-yd-Xd%*%as.matrix(betad[d,],nr=p)
  povinc.greg.var[d]<-sum(wd*(wd-1)*(gd*ed)^2)
}

# CVs des estimateurs GREG
povinc.greg.cv<-100*sqrt(povinc.greg.var)/povinc.greg
```

Nous représentons les valeurs des estimateurs GREG (en ordonnée) et celles du HT (en abscisse), ainsi que leurs variances (ou carrés des erreurs d'échantillonnage):

```
M<-max(povinc.dir,povinc.greg)
m<-min(povinc.dir,povinc.greg)
plot(povinc.dir,povinc.greg,ylim=c(m,M),xlim=c(m,M),xlab="HT",ylab="GREG")
abline(a=0,b=1)
M<-max(povinc.dir.var,povinc.greg.var)
m<-min(povinc.dir.var,povinc.greg.var)
plot(povinc.dir.var, povinc.greg.var, ylim=c(m,M), xlim=c(m,M), xlab="Var(HT)", ylab="Var(GREG)")
abline(a=0,b=1)
```

Figure 2
Estimateurs GREG de l'incidence de la pauvreté pour les provinces versus estimateurs HT (à droite), et variances estimées des estimateurs GREG versus estimateurs HT (à gauche)
(En proportions)



Source: Calculs de l'auteur.

Nous pouvons voir que les estimateurs GREG sont proches des estimateurs HT, mais que leurs variances estimées sont légèrement plus petites. Ce gain d'efficacité provient de l'utilisation d'information auxiliaire.

IV. Méthodes indirectes de base pour la désagrégation des données sur la pauvreté

Un estimateur indirect d'un indicateur relatif à un domaine spécifique est un estimateur qui utilise de l'information provenant d'autres domaines, en faisant l'hypothèse d'une certaine homogénéité entre ceux-ci. L'utilisation, dans le processus d'estimation, d'une quantité plus importante d'information que les seules données collectées sur le domaine conduit souvent à une diminution de l'erreur d'échantillonnage (et donc à une amélioration de l'efficacité de l'estimateur). Tout d'abord, nous présentons les estimateurs synthétiques. Un estimateur synthétique considère que des domaines sont homogènes au sens où ils ont des paramètres communs, sans offrir la possibilité d'une hétérogénéité entre eux. Ces estimateurs sont donc fondés sur des hypothèses fortes, et donc peu réalistes dans la pratique, et de ce fait ont des biais importants. En dépit de ce problème de biais potentiels, ils sont intégrés dans ce document, dans le but de présenter l'idée intuitive sous-jacente à l'estimation sur petits domaines, qui est d'emprunter de l'information aux autres domaines afin d'améliorer l'efficacité des estimateurs sur un domaine considéré.

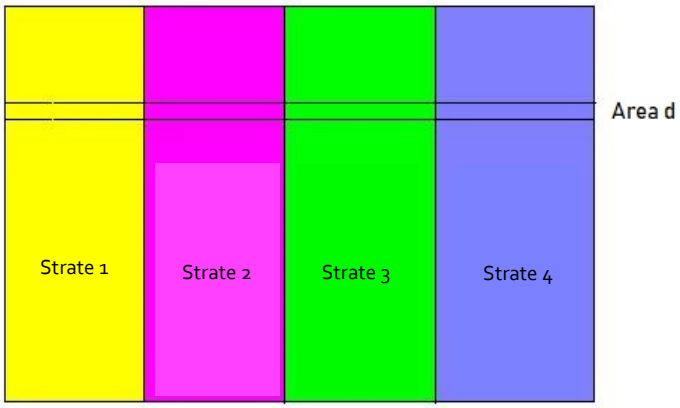
A. L'estimateur synthétique post-stratifié

Rappelons ici une nouvelle fois que cet estimateur est très rarement utilisé dans la pratique de l'estimation sur petits domaines, car fondé sur des hypothèses irréalistes; néanmoins, il est présenté dans ce chapitre car il donne une illustration simple du principe d'utilisation d'une information plus large.

Il existe une variable qualitative liée à la variable Y_{di}. Cette variable qualitative a J modalités possibles, qui divisent la population U en J groupes, $U^1, \dots, U^J$ de tailles $N^1, \dots, N^J$, appelés post-strates, celles-ci recoupant les domaines. De ce fait, le domaine U_d de la population est également divisé en J

« morceaux » de post-strates, $U_d^1, \dots, U_d^J$ de tailles $N_d^1, \dots, N_d^J$ et avec des valeurs moyennes $\bar{Y}_d^1, \dots, \bar{Y}_d^J$, où $\bar{Y}_d^j = \sum_{i \in U_d^j} Y_{di}/N_d^j, j = 1, \dots, J$ (voir Figure 3). Afin de simplifier, on utilisera ici le terme de strates pour les post-strates dans la figure 3 et dans la suite.

Figure 3
Population divisée en 4 post-strates et domaine d



Source: Auteur.

Etant donné que les moyennes sont des indicateurs additifs, nous pouvons les décomposer sur les J strates, de la façon suivante :

$$\bar{Y}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} Y_{di} = \frac{1}{N_d} \sum_{j=1}^{J} N_d^j \, \bar{Y}_d^j. \tag{11}$$

Il est maintenant supposé que les individus de chaque strate ont un comportement homogène, indépendamment du domaine auquel ils appartiennent, plus précisément, on fait l'hypothèse

$$\bar{Y}_d^j = \bar{Y}^j, \quad j = 1, \dots, J, \tag{12}$$

où $\bar{Y}^j = \sum_{i \in U^j} Y_{di} / N^j$ est la moyenne sur la strate j. Nous pouvons alors tenir compte de cette homogénéité au sein de la strate pour estimer la moyenne de chaque domaine en estimant la moyenne sur chaque strate (qui utilise une taille d'échantillon importante). En d'autres termes, en substituant (12) dans (11), nous obtenons

$$\bar{Y}_d = \frac{1}{N_d} \sum_{j=1}^{J} N_d^j \, \bar{Y}^j. \tag{13}$$

L'estimateur synthétique post-stratifié (PS-SYN) de $\bar{Y}_d$ est obtenu en estimant les moyennes de chaque strate dans (13) grâce à l'estimateur de Hájek:

$$\hat{\bar{Y}}_d^{PS-SYN} = \frac{1}{N_d} \sum_{j=1}^{J} N_d^j \, \hat{\bar{Y}}^{j,HA}.$$

Le nombre de strates J est censé être petit, et la taille d'échantillon dans chaque strate suffisamment importante. De ce fait, les estimateurs directs $\hat{\bar{Y}}^{j,HA}$ des moyennes dans la strate $\bar{Y}^j$ ont une variance petite. Ce qui signifie que quand on estime la moyenne du domaine d avec les estimateurs pour les strates $\hat{\bar{Y}}^{j,HA}$, la variance sera également petite.

L'homogénéité au sein de chaque strate est donc utilisée pour améliorer l'efficacité de l'estimateur du domaine d grâce à l'utilisation de l'ensemble des données de l'échantillon. Mais l'hypothèse d'homogénéité au sein de chaque strate (12) est irréaliste, et l'estimateur synthétique post-stratifié peut avoir un biais considérable.

Etant donné que le biais de ces estimateurs, plutôt que leur variance, est non négligeable et que cette dernière pourrait donc donner une image faussée de la qualité de l'estimateur, il est intéressant de considérer leur erreur quadratique moyenne (EQM), qui prend en compte les deux. Pour les estimateurs synthétiques, de manière générale, un estimateur de l'EQM (MSE en anglais : mean squared error) sous le plan peut s'écrire comme

$$\widehat{\mathrm{MSE}}_{\pi}(\hat{\bar{Y}}_d^{SYN}) = (\hat{\bar{Y}}_d^{SYN} - \hat{\bar{Y}}_d^{DIR})^2 - \widehat{\mathrm{var}}_{\pi}(\hat{\bar{Y}}_d^{DIR}),$$

(voir Rao et Molina (2015), p.44). Cet estimateur est très instable car il dépend de l'estimateur direct sur le domaine considéré. Des estimateurs plus stables de l'EQM ont été proposés, basés sur l'idée de « moyenner » sur l'ensemble des domaines, mais les résultats obtenus ne sont pas spécifiques à chaque domaine ; c'est-à-dire que la même valeur de l'EQM serait affectée à tous les domaines. Il n'existe pas, à notre connaissance, d'estimateurs de l'EQM pour les estimateurs synthétiques qui sont à la fois stables et spécifiques à chaque domaine. Cette caractéristique constitue un inconvénient des estimateurs synthétiques.

Utiliser l'estimateur PS-SYN pour un indicateur FGT serait, en principe, faisable en raison de l'additivité de ces indicateurs. Cependant, l'estimateur reposerait alors sur l'hypothèse (irréaliste) que l'indicateur FGT reste constant au sein de chaque strate, c'est-à-dire

$$F_{\alpha d}^{j} = F_{\alpha}^{j}, \quad j = 1, \dots, J,$$

si F_{α}^{j} est l'indicateur FGT dans la strate j. C'est pourquoi cet estimateur serait plus approprié pour estimer des moyennes ou des totaux de variables continues.

Les caractéristiques de ces estimateurs peuvent être résumées comme suit:

Indicateurs cibles: moyennes/totaux de la variable d'intérêt

Données requises:

- Poids de sondage w_{di} pour tous les individus de l'échantillon.
- Taille de la population du domaine, N_d, et tailles des populations des croisements domaine - strate, $N_d^j, j = 1, \dots, J$.
- Une variable qualitative (ou une combinaison de plusieurs) observée dans la même enquête que la variable d'intérêt, et qui lui est liée.

Avantages:

- Si la strate dispose de suffisamment d'observations dans l'échantillon, la variance peut être considérablement réduite par rapport à celle obtenue pour l'estimateur direct.

Inconvénients:

- Il n'est pas facile de trouver des estimateurs stables de l'EQM sous le dispositif spécifié.
- Dans la pratique, l'hypothèse d'homogénéité relative aux variables Y_{di} n'st pas réaliste. Si elle n'est pas vérifiée, les estimateurs résultants peuvent avoir un biais considérable, et de ce fait produire des résultats inappropriés. Par ailleurs, si on estime l'erreur d'échantillonnage, on trouve des petites valeurs. Et il est rarement possible d'estimer correctement le biais.

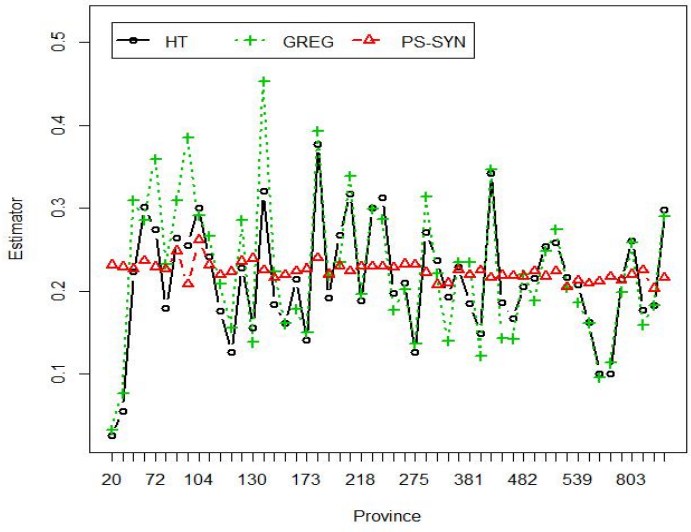
C'est pourquoi, en l'absence d'une méthode d'estimation du biais, les estimateurs pourraient apparaitre comme étant de bonne qualité, alors qu'il est peu probable que ce soit le cas.

Exemple 3. Estimateurs synthétiques post-stratifiés de l'incidence de la pauvreté sous R. Si l'on prolonge l'exemple 2, nous pouvons montrer comment calculer des estimateurs synthétiques post-stratifiés de l'incidence de la pauvreté pour des provinces, en utilisant le niveau d'éducation (variable educ) pour les post-strates.

Dans l'exemple 2, nous avons chargé les tailles de population des provinces pour chaque niveau d'éducation (base sizeprovedu). Ces tailles vont se trouver dans un objet "data frame" où les noms des colonnes doivent correspondre aux codes utilisés pour les catégories de la variable servant à post-stratifier (educ). Nous ajoutons donc les noms de colonnes au data frame avec les tailles de populations. Ensuite, nous appelons la fonction pssynt(), qui calcule les estimateurs post-stratifés pour l'incidence de la pauvreté (valeurs moyennes de la variable poor) en utilisant la variable educ et nous stockons les valeurs estimées:

```
colnames(sizeprovedu) <- c("provlab","prov","0","1","2","3")
povinc.psedu.res<-pssynt(y=poor,sweight=weight,ps=educ,domsizebyps=sizeprovedu[,-1])
povinc.psedu<-povinc.psedu.res$PsSynthetic
```

Figure 4
Estimateurs HT, GREG and PS-SYN de l'incidence de la pauvreté pour chaque province
(En proportions)



Source: Calculs de l'auteur.

Enfin, nous comparons sur un graphique les résultats avec ceux obtenus pour les estimateurs directs HT et GREG pour chaque province:

```
o<-order(nd)
M<-max(povinc.psedu,povinc.dir,povinc.greg)
m<-min(povinc.psedu,povinc.dir,povinc.greg)
k<-6
plot(1:D,povinc.dir[o],type="n",ylim=c(m,M+(M-m)/k),
     xlab="Province",ylab="Estimator",xaxt="n")
points(1:D,povinc.dir[o],type="b",col=1,lty=1,pch=1,lwd=2)
```

```
points(1:D,povinc.greg[o],type="b",col=3,lty=3,pch=3,lwd=2)
points(1:D,povinc.psedu[o],type="b",col=2,lty=2,pch=2,lwd=2)
 axis(1, at=1:D, labels=nd[o])
legend(1,M+(M-m)/k,legend=c("HT","GREG","PS-SYN"),ncol=3,
       col=c(1,3,2),lwd=rep(2,3), lty=c(1,3,2),pch=c(1,3,2))
```

Les résultats sont présentés dans la figure 4. Nous pouvons voir que les résultats produits par les estimateurs synthétiques post-stratifiés sont trop similaires entre les provinces, car ils supposent une homogénéité des individus ayant le même niveau d'éducation, indépendamment de la province où ils sont localisés. Cette hypothèse est très peu réaliste.

B. Estimateur synthétique de type régression au niveau domaine

Les estimateurs synthétiques de type régression supposent un modèle de régression linéaire situé soit au niveau domaine, soit au niveau individuel, selon l'information auxiliaire qui est disponible. Nous commençons par le cas où l'information auxiliaire n'est disponible qu'au niveau domaine. $\boldsymbol{x}_d$ désigne le vecteur des p variables auxiliaires disponibles au niveau du domaine (par exemple le vecteur des moyennes $\bar{\boldsymbol{X}}_d$ des p variables auxiliaires). Il est supposé que l'indicateur à estimer δ_d (par exemple, la moyenne sur le domaine, $\delta_d = \bar{Y}_d$) est lié, pour l'ensemble des domaines, aux données agrégées $\boldsymbol{x}_d$ par un modèle de régression linéaire. Puisque les vraies valeurs de l'indicateur dans les domaines ne sont pas disponibles (ce sont les paramètres cibles), on utilise à leur place les estimateurs directs $\hat{\delta}_d$, $d = 1, \dots, D$. Le modèle au niveau domaine suppose donc que

$$\hat{\delta}_d = \boldsymbol{x}_d{}'\boldsymbol{\alpha} + \varepsilon_d, \quad d = 1, \dots, D, \tag{14}$$

où les termes d'erreur ε_d sont supposés indépendants, avec une espérance nulle et une variance connue ψ_d , $d = 1, \dots, D$. Notons que puisque $\boldsymbol{x}_d$ est la valeur sur la population et a de ce fait une variance nulle, ψ_d est la variance de l'estimateur $\hat{\delta}_d$, donc $\psi_d = \mathrm{var}(\hat{\delta}_d)$. En pratique, ces variances sont estimées avec les données individuelles de l'enquête. L'estimateur synthétique de type régression (REG1-SYN) pour l'indicateur sur le domaine d s'écrit alors comme la valeur prédite par le modèle, c'est-à-dire que si $\widehat{\boldsymbol{\alpha}} = (\sum_{d=1}^{D} \psi_d^{-1} \boldsymbol{x}_d \boldsymbol{x}_d{}')^{-1} \sum_{d=1}^{D} \psi_d^{-1} \boldsymbol{x}_d \hat{\delta}_d$ est l'estimateur de $\boldsymbol{\alpha}$ obtenu par les moindres carrés pondérés, l'estimateur REG1-SYN de δ_d s'écrit

$$\hat{\delta}_d^{REG1-SYN} = \boldsymbol{x}_d{}'\widehat{\boldsymbol{\alpha}}.$$

Dans le modèle (14), ε_d est l'erreur due au fait que l'on utilise un estimateur direct $\hat{\delta}_d$ à la place de la vraie valeur de l'indicateur δ_d, celle-ci étant inconnue, et la vraie valeur δ_d est supposée être exactement égale au terme de la régression, $\delta_d = \boldsymbol{x}_d{}'\boldsymbol{\alpha}$, considérant ainsi qu'il n'y a aucune hétérogénéité possible entre les indicateurs des différents domaines au regard de cette régression. Les modèles de ce type, comme (14), qui n'incorporent pas d'effets d'hétérogénéité, sont appelés "modèles synthétiques". De ce fait, le biais sous le plan de $\hat{\delta}_d^{REG1-SYN}$ pour une valeur $\boldsymbol{\alpha}$ déterminée s'exprime comme $\boldsymbol{x}_d{}'\boldsymbol{\alpha} - \delta_d$, et il ne dépend pas de la taille de l'échantillon dans le domaine n_d; c'est pourquoi le biais ne diminue pas quand la taille de l'échantillon dans le domaine augmente.

Un avantage de ces estimateurs basés sur un modèle est qu'il permettent de produire des estimations dans les domaines où il n'y a pas eu d'unités échantillonnées, si l'information auxiliaire y est disponible. Pour un domaine d avec $n_d = 0$, si nous connaissons la valeur $\boldsymbol{x}_d$, l'estimateur synthétique de δ_d vaut de la même façon $\hat{\delta}_d^{REG1-SYN} = \boldsymbol{x}_d{}'\widehat{\boldsymbol{\alpha}}$.

Pour estimer l'indicateur FGT de pauvreté d'ordre α, $\delta_d = F_{\alpha d}$, si l'on utilise cette procédure, nous avons besoin de variables auxiliaires qui vérifient le modèle au niveau domaine

$$\hat{F}_{\alpha d} = \boldsymbol{x}_d{}'\boldsymbol{\alpha} + \varepsilon_d, \quad d = 1, \dots, D, \tag{15}$$

si $\psi_d = \mathrm{var}(\hat{F}_{\alpha d})$, $d = 1, \dots, D$ sont connues. L'estimateur synthétique de type régression pour l'indicateur FGT dans le domaine d, $F_{\alpha d}$, s'écrit alors comme

$$\hat{F}_{\alpha d}^{REG1-SYN} = \boldsymbol{x}_d{}'\widehat{\boldsymbol{\alpha}},$$

où, dans ce cas, $\widehat{\boldsymbol{\alpha}} = (\sum_{d=1}^{D} \psi_d^{-1} \boldsymbol{x}_d \boldsymbol{x}_d{}')^{-1} \sum_{d=1}^{D} \psi_d^{-1} \boldsymbol{x}_d \hat{F}_{\alpha d}$.

Le modèle (14) supposé par l'estimateur REG1-SYN lie ainsi tous les domaines au travers du paramètre commun de régression $\boldsymbol{\alpha}$. Quand on estime ce paramètre commun avec les estimateurs directs sur l'ensemble des domaines $\hat{\delta}_d$, on obtient un estimateur qui a une variance bien plus petite que celle des estimateurs directs. Cependant, ce modèle n'incorpore pas d'hétérogénéité entre les domaines, en-dehors de celle expliquée par les variables auxiliaires considérées. En pratique, il est difficile d'avoir des données auxiliaires qui expliquent complètement les variations des indicateurs δ_d dans les domaines pour lesquels on veut produire des estimations. C'est pourquoi le modèle synthétique (14) pourrait ne pas être représentatif de beaucoup des situations que l'on rencontre en pratique, et conduire à des biais dans ces cas. Notons de plus que, dans le cas le plus favorable où on connait le vrai modèle (et la vraie valeur de $\boldsymbol{\alpha}$), l'estimateur REG1-SYN serait $\boldsymbol{x}_d{}'\boldsymbol{\alpha}$, avec comme conséquence que les données recueillies via l'enquête pour la variable d'intérêt et pour le domaine considéré ne seraient pas utilisées. Ce qui pourrait être vu comme du gaspillage pour les domaines où la taille de l'échantillon est grande. De plus, l'estimateur obtenu peut être très différent de l'estimateur direct, alors que ce dernier est fiable pour les domaines en question. Ceci constitue un inconvénient majeur des estimateurs synthétiques (ou modèles). Mais par ailleurs, comme il a été indiqué dans l'introduction, comme ils sont potentiellement biaisés sous le plan, leur qualité devrait être évaluée en termes d'EQM, plutôt que de variance (qui sera sans doute faible, laissant penser, de manière erronée, que l'estimateur est de bonne qualité); mais il n'existe pas d'estimateurs connus de l'EQM sous le plan qui à la fois sont stables et produisent des résultats différents pour chaque domaine.

Les caractéristiques des ces estimateurs peuvent être résumées comme suit :

Indicateurs cibles: paramètres généraux.

Données requises:

- Données agrégées (par exemple moyennes sur des populations) des p variables auxiliaires considérées dans les domaines, $\boldsymbol{x}_d$, $d = 1, \dots, D$.

Avantages:

- La variance peut être considérablement réduite par rapport à celle de l'estimateur direct.
- On peut produire des estimations pour les domaines non échantillonnés.

Inconvénients:

- Le modèle de régression synthétique considéré n'est pas représentatif des cas où l'hétérogénéité entre domaines n'est pas expliquée par les seules variables auxiliaires prises en compte. Dans ces cas, les estimateurs produits peuvent avoir des biais substantiels.
- Il est nécessaire d'analyser le modèle de manière appofondie (par exemple sur les résidus), car le biais de ces estimateurs dépend de la bonne qualité de l'ajustement du modèle. En particulier, il est très important de vérifier s'il y a une effet "domaine", car le modèle ne le prend pas en compte.
- Si le modèle est connu, les données collectées pour la variable d'intérêt et sur le domaine ne sont pas utilisées.

- L'estimateur reste différent de l'estimateur direct quand la taille d'échantillon augmente.
- Il n'existe pas d'estimateurs connus de l'EQM sous le plan qui à la fois sont stables et produisent des résultats différents pour chaque domaine.
- Ils demandent des ajustements afin de pouvoir assurer la propriété de réconciliation des données, à savoir que la somme des totaux estimés pour différents domaines retrouve la valeur totale.

C. Estimateur synthétique de type régression au niveau individuel

Nous considérons maintenant que des données au niveau individu (ou microdonnées) sont disponibles, pour p variables auxiliaires de l'enquête, $\boldsymbol{x}_{di}, i \in s_d, d = 1, \dots, D$. Dans ce cas, un estimateur synthétique de type régression peut être obtenu, pour l'indicateur d'intérêt, en posant un modèle de régression linéaire au niveau individuel pour Y_{di}. $\boldsymbol{y}_d = (Y_{d1}, \dots, Y_{dN_d})'$ est le vecteur des valeurs de la variable en question pour les individus du domaine d. L'indicateur à estimer pour le domaine d est une fonction de ce vecteur, par exemple, $\delta_d = \delta_d(\boldsymbol{y}_d)$. Le modèle synthétique de régression de base considère que les variables Y_{di} suivent, pour tous les individus de la population, le modèle de régression linéaire

$$Y_{di} = \boldsymbol{x}_{di}'\boldsymbol{\beta} + \varepsilon_{di}, \quad i = 1, \dots, N_d, d = 1, \dots, D, \tag{16}$$

où les erreurs ε_{di} sont indépendantes, d'espérance nulle et de variance $\sigma^2 k_{di}^2$, où les k_{di} sont des constantes connues qui représentent l'hétéroscédasticité posible au sein du modèle ($k_{di} = 1$ pour tout i et d s'il n'y a pas d'hétéroscédasticité). En estimant $\boldsymbol{\beta}$ grâce à l'estimateur des moindres carrés pondérés $\widehat{\boldsymbol{\beta}} = (\sum_{d=1}^{D}\sum_{i\in s_d} a_{di}\,\boldsymbol{x}_{di}\boldsymbol{x}_{di}')^{-1}\sum_{d=1}^{D}\sum_{i\in s_d} a_{di}\,\boldsymbol{x}_{di}Y_{di}$, où $a_{di} = k_{di}^{-2}$, nous obtenons des valeurs prédites par le modèle pour chaque indivdu du domaine, $\hat{Y}_{di} = \boldsymbol{x}_{di}'\widehat{\boldsymbol{\beta}}, i = 1, \dots, N_d$. Le vecteur des valeurs prédites dans le domaine d vaut donc $\widehat{\boldsymbol{y}}_d = (\hat{Y}_{d1}, \dots, \hat{Y}_{dN_d})'$. En utilisant ce vecteur à la place des $\boldsymbol{y}_d$ pour calculer l'indicateur, nous obtenons l'estimateur sythétique de type régression de δ_d, soit

$$\hat{\delta}_d^{REG2-SYN} = \delta_d(\widehat{\boldsymbol{y}}_d).$$

Par exemple, pour la moyenne du domaine d, $\delta_d = \bar{Y}_d$, si $\bar{\boldsymbol{X}}_d$ est le vecteur des moyennes des p variables auxiliaires utilisées, l'estimateur synthétique pour le modèle (16) devient

$$\hat{\bar{Y}}_d^{REG2-SYN} = \bar{\boldsymbol{X}}_d{}'\widehat{\boldsymbol{\beta}}.$$

Pour un domaine non échantillonné, l'estimateur est obtenu de la même manière. Pour une valeur $\boldsymbol{\beta}$ connue, le biais sous le plan de l'estimateur de la moyenne est $\bar{\boldsymbol{X}}_d{}'\boldsymbol{\beta} - \bar{Y}_d$, et il ne dépend pas de la taille de l'échantillon dans le domaine n_d; de ce fait, ce biais ne décroît pas quand la taille de l'échantillon augmente.

A nouveau, si nous cherchons à estimer les indicateurs FGT de pauvreté, nous devons trouver des variables $\boldsymbol{x}_{di}$ liées de façon linéaire à $F_{\alpha,di}$; par exemple, qui suivent le modèle

$$F_{\alpha,di} = \boldsymbol{x}_{di}'\boldsymbol{\beta} + \varepsilon_{di}, \quad i = 1, \dots, N_d, d = 1, \dots, D. \tag{17}$$

Cependant, trouver des variables reliées de façon linéaire aux indicateurs $F_{\alpha,di}$ est rarement réalisé en pratique. Il est plus approprié de supposer le modèle vérifié pour les variables d'intérêt, c'est-à-dire celles qui sont utilisées pour mesurer le pouvoir d'achat, E_{di} ou, mieux encore, pour une transformation bijective de ces variables, $T(E_{di})$ car, les E_{di} ayant en général une distribution très asymétrique, un modèle linéaire qui leur serait appliqué serait sans doute inapproprié. En pratique, il est classique d'utiliser une transformation logarithmique, à savoir que $Y_{di} = \log(E_{di} + c)$ est utilisée comme variable de réponse dans le modèle, avec $c > 0$ qui est une constante positive, rendant la distribution de Y_{di} approximativement normale. Cette constante peut être déterminée en ajustant le

modèle pour une série de valeurs de c dans l'étendue de E_{di}, et en prenant la valeur de c pour laquelle la mesure de l'asymétrie des résidus du modèle (par exemple, le coefficient d'asymétrie de Pearson) est aussi proche que possible de zéro.

Comme dans le cas des estimateurs synthétiques précédents, si l'hétérogénéité des Y_{di} dans les domaines n'est pas parfaitement expliquée par l'ensemble des variables auxiliaires (donc, si le modèle synthétique posé n'est pas valable), les estimateurs seront biaisés. Mais leur variance sera faible puisque le coefficient de régression est estimé grâce à l'échantillon total, qui est en général de grande taille. De ce fait, l'estimateur synthétique de type régression aura une petite erreur d'échantillonnage. Ces estimateurs nécessitent donc une étude de la qualité de l'ajustement du modèle posé, afin d'éviter d'avoir de larges biais. Et de nouveau, à supposer que l'on connaisse le modèle de façon parfaite, ces estimateurs n'utiliseraient que les données des variables auxiliaires, et non les données de la variable d'intérêt collectées dans le domaine considéré, et ils ne se rapprochent pas nécessairement des estimateurs directs pour les domaines qui ont une taille d'échantillon importante. De plus, il n'existe pas d'estimateurs fiables de l'EQM sous le dispositif spécifié qui produisent des valeurs différentes pour chaque domaine.

Les caractéristiques de ces estimateurs peuvent être résumées comme suit:

Indicateurs cibles: paramètres généraux.

Données requises:

- Données sur l'échantillon pour les p variables auxiliaires liées à la variable d'intérêt, collectées par la même enquête que celle où la variable d'intérêt est disponible.
- Pour des indicateurs de moyennes/totaux de la variable de réponse considérée dans le modèle, moyennes/totaux des p variables auxiliaires dans les domaines, $\bar{\boldsymbol{X}}_d$, $d = 1, \dots, D$. Pour des indicateurs non linéaires des variables de réponse du modèle, les valeurs des p variables auxiliaires sont nécessaires pour tous les individus (microdonnées) dans le domaine $\{\boldsymbol{x}_{di};\ i = 1, \dots, N_d, d = 1, \dots, D\}$.

Avantages:

- Il est possible de réduire de manière très nette la variance des estimateurs directs et des estimateurs obtenus à partir d'un modèle au niveau domaine.
- L'estimateur peut être utilisé dans des domaines non échantillonnés.

Inconvénients:

- Le modèle synthétique de régression considéré n'est pas approprié pour les cas où l'hétérogénéité entre domaines n'est pas expliquée totalement par les variables auxiliaires disponibles. C'est pourquoi, dans certains cas, les estimateurs résultants peuvent avoir un biais important.
- Il est nécessaire d'analyser le modèle de manière approfondie (par exemple par l'intermédiaire des résidus), car le biais des estimateurs dépend de la qualité de l'ajustement du modèle. En particulier, il est très important de vérifier s'il y a un effet "domaine", car le modèle ne le prend pas en compte.
- Si le modèle était connu de façon parfaite, l'estimateur n'utiliserait pas des données de la variable d'intérêt pour le domaine.
- De la même manière, l'estimateur ne se rapproche pas de l'estimateur direct quand la taille d'échantillon augmente.

- Il n'existe pas d'estimateurs connus de l'EQM selon le plan qui sont stables et fournissent des valeurs différentes pour chaque domaine.
- Ils demandent des ajustements afin de pouvoir assurer la propriété de réconciliation des données, à savoir que la somme des totaux estimés pour différents domaines retrouve la valeur totale.

D. Les estimateurs composites

Comme indiqué dans les chapitres précédents, les estimateurs directs sont (au moins approximativement) sans biais sous le plan de sondage, mais peuvent avoir une variance importante pour les domaines où la taille d'échantillon est faible. A l'inverse, les estimateurs synthétiques ont une variance faible, mais peuvent avoir un biais très important selon le dispositif spécifié. Les estimateurs composites sont conçus pour diminuer la variance de l'estimateur direct, tout en acceptant une partie du biais de l'estimateur synthétique. L'objectif recherché est, simultanément, d'améliorer l'efficacité de l'estimateur direct et de réduire le biais de l'estimateur synthétique. Soit $\hat{\bar{Y}}_d^{DIR}$ un estimateur générique direct de $\bar{Y}_d$ et $\hat{\bar{Y}}_d^{SYN}$ un estimateur synthétique. Un estimateur composite de $\bar{Y}_d$ aura la forme suivante:

$$\hat{\bar{Y}}_d^{C} = \phi_d \hat{\bar{Y}}_d^{DIR} + (1-\phi_d)\hat{\bar{Y}}_d^{SYN}, \quad 0 \leq \phi_d \leq 1.$$

Le poids ϕ_d donné à l'estimateur direct peut être déterminé soit en minimisant une approximation de l'erreur quadratique moyenne (EQM) sous le plan de sondage, ce qui ne peut se faire que de manière approchée, soit en le fixant de manière arbitraire. Drew, Singh, et Choudhry (1982) ont proposé un poids ϕ_d qui dépend de la taille d'échantillon dans le domaine, ce qui conduit à l'estimateur SSD (en anglais sample-size dependent). Si l'on prend une valeur fixée a priori $\delta > 0$ (par défaut on peut prendre la valeur 1), le poids proposé sera de la forme

$$\phi_d = \begin{cases} 1, & \text{si} \quad \widehat{N}_d \geq \delta N_d; \\ \widehat{N}_d/(\delta N_d), & \text{si} \quad \widehat{N}_d < \delta N_d, \end{cases}$$

où $\widehat{N}_d = \sum_{i \in s_d} w_{di}$. Pour comprendre l'idée intuitive qui sous-tend cet estimateur, rappelons que, pour un sondage aléatoire simple (simple random sampling, SRS) dans la population (dans ce cas les tailles d'échantillon dans les domaines sont aléatoires), on obtient:

$$\widehat{N}_d = \sum_{i \in s_d} w_{di} = \sum_{i \in s_d} \frac{N}{n} = N\frac{n_d}{n}.$$

Et puisque $\widehat{N}_d$ est sans biais, son espérance sous le plan de sondage est égale à $NE_\pi(n_d)/n = N_d$, donc $E_\pi(n_d) = nN_d/N$ et de ce fait le poids proposé vaut

$$\phi_d = \begin{cases} 1 & \text{si} \quad n_d \geq \delta E_\pi(n_d); \\ n_d/\{\delta E_\pi(n_d)\} & \text{si} \quad n_d < \delta E_\pi(n_d). \end{cases}$$

Si l'on choisit $\delta = 1$, l'estimateur SSD donne un poids de 1 à l'estimateur direct si la taille de l'échantillon dans le domaine est supérieure ou égale à la taille attendue, et un poids inférieur à 1 dans le cas contraire. Mais un domaine donné peut avoir une petite taille d'échantillon n_d , qui excède quand même la taille attendue, ce qui donnerait un poids de 1 à l'estimateur direct, et ne conduirait pas à une amélioration de l'efficacité au final.

L'estimateur SSD a été utilisé dans l'enquête emploi canadienne afin d'obtenir des estimateurs pour les secteurs de recensement, en prenant $\delta = 2/3$ (voir Drew, Singh, et Choudhry (1982)). Cependant, pour la plupart des secteurs considérés, le poids de l'estimateur direct s'est avéré être $\phi_d = 1$; pour quelques-uns, le poids valait $\phi_d = 0.9$, mais en aucun cas le poids obtenu n'est descendu en-dessous de 0.8. De ce fait, le gain en efficacité, par rapport à l'estimateur direct, a été très limité.

Comme dans cet exemple, le problème de cet estimateur est qu'il tend à donner à l'estimateur direct un poids proche de 1 même si la taille d'échantillon dans le domaine est faible, avec pour conséquence aucune amélioration de l'efficacité (si l'on compare à l'estimateur direct). De plus, le poids ϕ_d ne prend pas en considération le fait que les domaines sont plus ou moins homogènes relativement au modèle considéré pour l'estimateur synthétique. Il est donc indépendant de la qualité de l'estimateur synthétique (ou de la qualité de l'ajustement du modèle synthétique) pour chaque domaine. De ce fait, ces estimateurs peuvent être vus comme trop simplifiés pour fournir une amélioration sensible, si on les compare aux estimateurs directs.

Comme indiqué précédemment, il est possible d'obtenir des estimateurs composites approximativement optimaux relativement au plan de sondage en prenant le poids ϕ_d qui minimise (approximativement) l'EQM de l'estimateur composite spécifié, $\mathrm{MSE}_\pi(\hat{\bar{Y}}_d^C)$. Si l'on considère que la covariance entre l'estimateur direct et l'estimateur synthétique est négligeable, et en minimisant

$$\mathrm{MSE}_\pi(\hat{\bar{Y}}_d^C) \approx \phi_d^2 \mathrm{var}_\pi(\hat{\bar{Y}}_d^{DIR}) + (1-\phi_d)^2 \mathrm{MSE}_\pi(\hat{\bar{Y}}_d^{SYN}),$$

le poids optimal vaut

$$\phi_d^* = \mathrm{MSE}_\pi(\hat{\bar{Y}}_d^{SYN}) / \{\mathrm{var}_\pi(\hat{\bar{Y}}_d^{DIR}) + \mathrm{MSE}_\pi(\hat{\bar{Y}}_d^{SYN})\}. \tag{18}$$

Un estimateur de $\mathrm{MSE}_\pi(\hat{\bar{Y}}_d^{SYN})$ est

$$\widehat{\mathrm{MSE}}_\pi(\hat{\bar{Y}}_d^{SYN}) = (\hat{\bar{Y}}_d^{SYN} - \hat{\bar{Y}}_d^{DIR})^2 - \widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d^{DIR}),$$

(voir Rao et Molina (2015, p.44)). En remplaçant cet estimateur dans le poids optimal ϕ_d^* donné en (18),nous obtenons un estimateur de ce poids optimal, qui s'écrit

$$\hat{\phi}_d^* = \widehat{\mathrm{MSE}}_\pi(\hat{\bar{Y}}_d^{SYN}) / (\hat{\bar{Y}}_d^{SYN} - \hat{\bar{Y}}_d^{DIR})^2 = 1 - \widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d^{DIR}) / (\hat{\bar{Y}}_d^{SYN} - \hat{\bar{Y}}_d^{DIR})^2.$$

Nous pouvons voir que ce poids dépend de l'estimateur direct $\hat{\bar{Y}}_d^{DIR}$, qui est très volatile. Ceci veut dire que le poids optimal estimé $\hat{\phi}_d^*$ est également très volatile. Un poids estimé plus stable peut être obtenu en moyennant sur l'ensemble des domaines, comme suit:

$$\begin{aligned}\hat{\phi}^* &= \sum_{\ell=1}^{D} \widehat{\mathrm{MSE}}_\pi(\hat{\bar{Y}}_\ell^{SYN}) / \sum_{\ell=1}^{D} (\hat{\bar{Y}}_\ell^{SYN} - \hat{\bar{Y}}_\ell^{DIR})^2 \\ &= 1 - \left\{ \sum_{\ell=1}^{D} \widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_\ell^{DIR}) / \sum_{\ell=1}^{D} (\hat{\bar{Y}}_\ell^{SYN} - \hat{\bar{Y}}_\ell^{DIR})^2 \right\}\end{aligned}$$

Le poids résultant, $\hat{\phi}^*$, est très stable, mais ne dépend pas du domaine d; ce qui signifie qu'il est constant pour tous les domaines, et ne dépend pas de leur taille d'échantillon. En raison de ces inconvénients, les estimateurs composites optimaux sont probablement moins utilisés en pratique que les estimateurs basés sur un modèle qui seront présentés dans le chapitre suivant.

Les estimateurs composites sont intéressants, car ils cherchent un compromis entre biais et variance. Néanmoins, dans les chapitres suivants, nous verrons que des estimateurs composites peuvent être obtenus de manière plus efficace à partir de modèles de régression qui prennent en compte l'hétérogénéité entre domaines. Ces estimateurs composites seront optimaux relativement aux distributions générées par les modèles posés, c'est pourquoi ils sont appelés estimateurs basés sur un modèle. Pour ces estimateurs, les poids dépendent de la taille d'échantillon du domaine et de la qualité de l'ajustement du modèle synthétique, avec un poids plus grand accordé à l'estimateur direct si le modèle synthétique est de mauvaise qualité (les variables auxiliaires n'apportent pas une information utile, ou les domaines sont très hétérogènes) ou si la taille d'échantillon du domaine est importante, et une valeur du poids plus importante est accordée à l'estimateur synthétique quand la taille d'échantillon

décroît ou si le modèle a de meilleures qualités prédictives. De ce fait, les estimateurs basés sur un modèle « surpassent » ces estimateurs composites simples.

Nous résumons dans ce qui suit les caractéristiques de l'estimateur SSD, en tant que représentant le plus commun des estimateurs composites:

Indicateurs cibles: paramètres additifs.

Données requises:

- Poids de sondage w_{di} pour les individus échantillonnés dans le domaine d.
- Taille de la population du domaine, N_d, si l'estimateur HT de la moyenne, ou l'estimateur de Hájek du total, est utilisé.

Avantages:

- Ils sont conçus pour réduire à la fois le biais de l'estimateur synthétique et la variance de l'estimateur direct. Ils ne peuvent pas être moins efficaces que l'estimateur direct et le biais ne peut pas être plus grand que celui de l'estimateur synthétique.

Inconvénients:

- Pour un domaine de petite taille, tant que cette taille n'est pas plus petite que celle attendue, aucune information n'est "empruntée" aux autres domaines par l'intermédiaire de l'estimateur synthétique. De ce fait, il n'y aura pas de gain d'efficacité par rapport à l'estimateur direct.
- Le poids accordé à l'estimateur synthétique ne dépend pas de la qualité du lien entre la variable d'intérêt et les variables explicatives, donc du bon ajustement du modèle.
- Ces estimateurs ne peuvent pas être calculés pour des domaines non échantillonnés (c'est-à-dire ceux où la taille d'échantillon n_d est égale à zéro).
- Il n'existe pas d'estimateurs connus et stables de l'EQM selon le dispositif spécifié qui soient différents selon les domaines.
- Ils demandent des ajustements afin de pouvoir assurer la propriété de réconciliation des données, à savoir que la somme des totaux estimés pour différents domaines retrouve la valeur totale.

Exemple 4. Estimateurs composites de l'incidence de la pauvreté, avec R. Dans le prolongement des exemples précédents, nous montrons maintenant comment produire des estimateurs composites de l'incidence de la pauvreté pour les provinces, en utilisant les estimateurs directs HT de l'exemple 1 et les estimateurs synthétiques post-stratifiés de l'exemple 3. Pour cela, nous appelons la fonction ssd() en utilisant la valeur par défaut du paramètre delta (delta=1) et sauvegardons les résultats:

```
povinc.ssd.res<-ssd(dom=prov,sweight=weight,domsize=sizeprov[,c(2,3)],
 direct=povinc.dir.res[,c("Domain","Direct")],synthetic=povinc.psedu.res)
povinc.ssd<-povinc.ssd.res$ssd
```

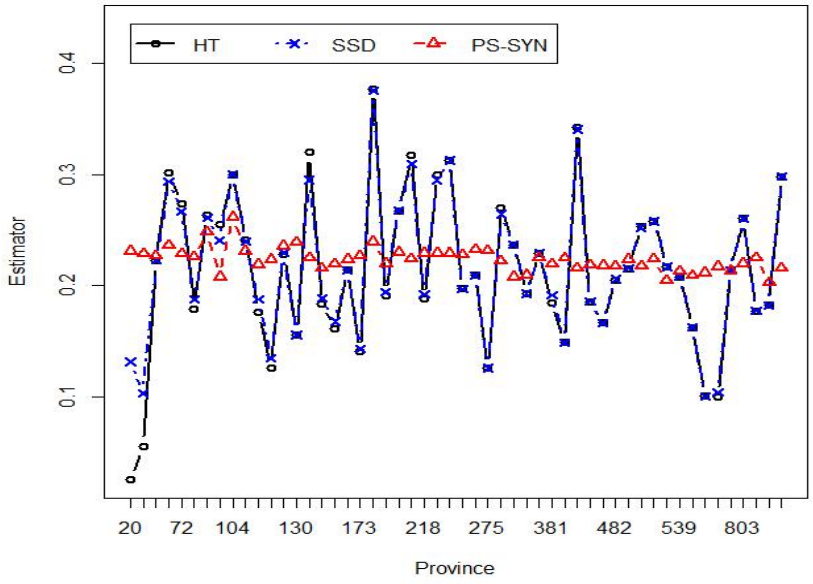
Nous analysons les poids donnés par l'estimateur SSD à l'estimateur direct pour chaque province par l'intermédiaire de statistiques descriptives de ces poids:

```
summary(povinc.ssd.res$CompWeight)
```

Le résultat est:

```
   Min.   1st Qu.   Median     Mean  3rd Qu.     Max.
 0.4846   0.8800   0.9779   0.9224   1.0000   1.0000
```

Figure 5
Estimateurs HT, PS-SYN et SSD de l'incidence de la pauvreté pour chaque province
(En proportions)



Source: Calculs de l'auteur.

Nous pouvons voir que les estimateurs directs sont affectés d'un poids égal à 1 pour au moins un quart des provinces. Pour celles-ci, aucune information n'est "empruntée" aux autres provinces. Par ailleurs, pour cet estimateur le poids ne dépend pas de la variable d'intérêt. Si, par exemple, on cherche à estimer le revenu moyen, nous obtenons exactement le même poids. Et si l'on compare sur un graphique les estimations SSD obtenues avec les estimations directes HT et les estimations synthétiques post-stratifiées (figure 5), nous pouvons voir qu'elles sont très proches des estimations directes HT. Sur ce graphique, les provinces (en abscisse) sont ordonnées des plus petites tailles d'échantillon aux plus grandes, et leurs tailles d'échantillon sont indiquées sur l'axe. Le code R servant à obtenir ces résultats est le suivant:

```
o<-order(nd)
k<-2
M<-max(povinc.psedu,povinc.dir,povinc.ssd)
m<-min(povinc.psedu,povinc.dir,povinc.ssd)
plot(1:D,povinc.dir[o],type="n",ylim=c(m,M+(M-m)/k),xlab="Province", ylab="Estimator",xaxt="n")
points(1:D,povinc.dir[o],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:D,povinc.greg[o],type="b",col=3,lty=3,pch=3,lwd=2)
points(1:D,povinc.ssd[o],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:D,povinc.psedu[o],type="b",col=2,lty=2,pch=2,lwd=2)
axis(1, at=1:D, labels=nd[o])
legend(1,M+(M-m)/k,legend=c("HT","GREG","SSD","PS-SYN"),ncol=4,col=c(1,3,4,2),
lwd=rep(2,3),lty=c(1,3,4,2),pch=c(1,3,4,2))
```

V. Méthodes indirectes fondées sur des modèles

Les estimateurs sur petits domaines fondés sur des modèles entrent dans le groupe des estimateurs indirects puisqu'ils empruntent des informations à d'autres domaines. Cependant, ils sont un peu plus sophistiqués que les estimateurs indirects de base discutés au chapitre IV, dans la mesure où ils incorporent l'hétérogénéité entre les domaines qui n'est pas expliquée par les variables auxiliaires considérées. On réalise ceci en incorporant des effets aléatoires additifs sur les domaines dans le modèle de régression considéré. Nous verrons que ces effets aléatoires confèrent une très bonne propriété aux estimateurs fondés sur un modèle linéaire, à savoir qu'ils peuvent être écrits comme des estimateurs composites qui conduisent à un estimateur direct dans les zones ayant une taille d'échantillon suffisante. Il est rare de disposer de toutes les variables qui expliquent entièrement l'hétérogénéité entre les domaines pour notre variable d'intérêt. Par conséquent, ces modèles sont nettement plus réalistes que les modèles synthétiques, ce qui permet d'obtenir des estimateurs présentant un biais plus faible sous le plan d'échantillonnage.

A. Estimateur EBLUP fondé sur le modèle de Fay-Herriot

Le modèle de Fay-Herriot (FH) est un modèle classique au niveau régional qui a été introduit par Fay et Herriot (1979) pour estimer le revenu par habitant dans de petites régions des États-Unis. Ce modèle est actuellement utilisé par le *Bureau of Census* des États-Unis dans le cadre du programme "Small Area Income and Poverty Estimates (SAIPE)"[2] pour estimer les proportions d'enfants pauvres en âge scolaire dans les comtés et les districts scolaires (pour plus de détails, voir Bell (1997) ou http://www.census.gov/hhes/www/saipe). Ce modèle a également été utilisé au Chili pour estimer les taux d'incidence de la pauvreté dans les communes chiliennes (voir Casas-Cordero Valencia, Encina et Lahiri (2015)) et en

[2] Estimations du revenu et de la pauvreté dans des petits domaines.

Espagne pour estimer l'incidence de la pauvreté et l'écart de pauvreté dans les provinces selon le genre (Molina et Morales, 2009).

Ce modèle relie les indicateurs d'intérêt pour tous les domaines δ_d, $d = 1, \dots, D$, en faisant l'hypothèse qu'ils varient en fonction d'un vecteur des valeurs de p variables auxiliaires $\boldsymbol{x}_d$ systématiquement pour tous les domaines, selon le modèle de regression linéaire

$$\delta_d = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d, \quad d = 1, \dots, D, \tag{19}$$

où $\boldsymbol{\beta}$ est le vecteur des coefficients communs à tous les domaines et u_d est le terme d'erreur, différent pour chaque domaine, s'interprétant comme l'effet aléatoire relatif au domaine d. Ces effets aléatoires u_d représentent l'hétérogénéité des indicateurs δ_d parmi les domaines, qui n'est pas due à (ou n'est pas expliquée par) les variables auxiliaires considérées. Dans le modèle le plus simple, de tels effets aléatoires u_d sont supposés indépendants et identiquement distribués (IID), avec une variance commune inconnue σ_u^2 ; on note ceci : $u_d \sim^{iid}(0, \sigma_u^2)$.

Puisque les vraies valeurs des indicateurs δ_d ne sont pas observables, le modèle (19) ne peut être estimé. Quand on utilise un estimateur direct $\hat{\delta}_d^{DIR}$ de δ_d,on doit considérer que cet estimateur a une erreur d'échantillonnage. Le modèle FH considère que cet estimateur $\hat{\delta}_d^{DIR}$ est sans biais sous le plan de sondage. Dans ce cas, on peut représenter l'erreur d'échantillonnage de cet estimateur en utilisant le modèle :

$$\hat{\delta}_d^{DIR} = \delta_d + e_d, \quad d = 1, \dots, D, \tag{20}$$

où e_d est l'erreur d'échantillonnage dans le domaine d. On fait l'hypothèse que les erreurs d'échantillonnage e_d sont indépendantes l'une de l'autre et sont aussi indépendantes des effets aléatoires sur les domaines, u_d, et sont d'espérance nulle et de variances connues ψ_d; i.e., $e_d \sim^{ind}(0, \psi_d)$. Dans la pratique, ces variances, $\psi_d = \mathrm{var}_\pi(\hat{\delta}_d^{DIR}|\delta_d)$, $d = 1, \dots, D$, sont estimées en utilisant les microdonnées issues de l'enquête. En combinant les modèles (19) and (20), on obtient le modèle linéaire mixte qui s'écrit

$$\hat{\delta}_d^{DIR} = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d + e_d, \quad d = 1, \dots, D. \tag{21}$$

En utilisant la méthode des multiplicateurs de Lagrange pour calculer l'estimateur linéaire en les données $\hat{\delta}_d^{DIR}$, $d = 1, \dots, D$, qui est sans biais sous le modèle (21), et qui minimise l'EQM sous le modèle, on obtient le meilleur prédicteur sans biais (BLUP)[3] de $\delta_d = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d$ sous le modèle. L'estimateur qui en résulte est obtenu en ajustant simplement le modèle mixte (21); i.e., l'estimateur BLUP sous le modèle FH a pour expression

$$\tilde{\delta}_d^{FH} = \boldsymbol{x}_d{}'\tilde{\boldsymbol{\beta}} + \tilde{u}_d, \tag{22}$$

où $\tilde{u}_d = \gamma_d(\hat{\delta}_d^{DIR} - \boldsymbol{x}_d{}'\tilde{\boldsymbol{\beta}})$ est l'estimateur BLUP de u_d, où $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d)$ et où $\tilde{\boldsymbol{\beta}}$ est l'estimateur des moindres carrés pondérés de $\boldsymbol{\beta}$ sous le modèle (21), qui s'exprime sous la forme :

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \gamma_d \, \boldsymbol{x}_d \boldsymbol{x}_d{}'\right)^{-1} \sum_{d=1}^{D} \gamma_d \, \boldsymbol{x}_d \hat{\delta}_d^{DIR}.$$

Notons que, en écrivant $\tilde{u}_d = \gamma_d(\hat{\delta}_d^{DIR} - \boldsymbol{x}_d{}'\tilde{\boldsymbol{\beta}})$ dans l'estimateur BLUP sous le modèle FH donné en (22), nous pouvons exprimer l'estimateur BLUP comme une combinaison linéaire convexe de l'estimateur direct et de l'estimateur synthétique par régression, i.e,

$$\tilde{\delta}_d^{FH} = \gamma_d \hat{\delta}_d^{DIR} + (1 - \gamma_d)\boldsymbol{x}_d{}'\tilde{\boldsymbol{\beta}}, \tag{23}$$

[3] En anglais : « best linear unbiased predictor». L'acronyme anglais sera conservé par la suite.

avec un poids pour l'estimateur direct exprimé par $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d) \in (0,1)$. Ce poids dépend de la taille de l'échantillon du domaine par l'intermédiaire de la variance ψ_d de l'estimateur direct et de la qualité de l'ajustement du modèle synthétique mesurée par σ_u^2 (en d'autres termes, l'hétérogénéité inexpliquée entre les domaines). Par conséquent, pour un domaine d dans lequel l'estimateur direct $\hat{\delta}_d^{DIR}$ est efficace en raison d'une taille d'échantillon suffisante ; c'est-à-dire avec une petite variance d'échantillonnage ψ_d par rapport à l'hétérogénéité inexpliquée σ_u^2, $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d)$ est proche de un et donc $\tilde{\delta}_d^{FH}$ donne plus de poids à l'estimateur direct. En revanche, dans les domaines d où l'estimateur direct manque de qualité en raison de la petite taille de l'échantillon, où sa variance d'échantillonnage ψ_d est plus grande que l'hétérogénéité inexpliquée σ_u^2, alors γ_d s'approche de zéro et l'on donne ainsi plus de poids à l'estimateur synthétique par la régression $\boldsymbol{x}_d{}'\widetilde{\boldsymbol{\beta}}$, qui utilise les données de tous les domaines pour estimer le paramètre commun $\boldsymbol{\beta}$. En d'autres termes, cet estimateur utilise des informations provenant des autres domaines au moyen de l'estimateur synthétique par la régression $\boldsymbol{x}_d{}'\widetilde{\boldsymbol{\beta}}$ lorsque cela est nécessaire, en fonction de l'efficacité de l'estimateur direct.

De plus, le fait que l'estimateur BLUP $\tilde{\delta}_d^{FH}$ se rapproche de l'estimateur direct lorsque la taille de l'échantillon du domaine est grande (ψ_d petit) est une propriété très souhaitable, puisque nous n'avons pas besoin de savoir quand un domaine est suffisamment "petit" pour utiliser cet estimateur à la place de l'estimateur direct, puisqu'il tend vers l'estimateur direct lorsque la taille de l'échantillon augmente, et qu'il améliore également l'estimateur direct dans les domaines ayant un petit échantillon. Par conséquent, en principe, cet estimateur peut être utilisé pour tous les domaines à partir du moment où il en existe un "petit" (s'il n'y en avait pas, il ne serait pas nécessaire de l'utiliser).

L'estimateur BLUP de δ_d dépend de la vraie valeur de la variance σ_u^2 des effets aléatoires u_d. En pratique, cette variance est inconnue et doit être estimée. Les méthodes d'estimation courantes sont le maximum de vraisemblance (ML) et le ML restreint/résiduel (REML). La méthode REML corrige l'estimateur de la variance σ_u^2 ou les degrés de liberté liés à l'estimation des coefficients de régression $\boldsymbol{\beta}$ et fournit ainsi un estimateur moins biaisé pour une taille d'échantillon finie n. Une méthode d'ajustement basée sur les moments, qui ne nécessite pas une distribution paramétrique pour obtenir la vraisemblance, est celle avancée par Fay et Herriot (1979), que nous appelons la méthode FH. Soit $\hat{\sigma}_u^2$ un estimateur convergent de σ_u^2 comme ceux obtenus par ces méthodes. En remplaçant σ_u^2 par $\hat{\sigma}_u^2$ en (22), nous obtenons l'estimateur BLUP empirique (EBLUP[4]) de δ_d,

$$\hat{\delta}_d^{FH} = \hat{\gamma}_d \hat{\delta}_d^{DIR} + (1 - \hat{\gamma}_d)\boldsymbol{x}_d{}'\widehat{\boldsymbol{\beta}}, \qquad (24)$$

où $\hat{\gamma}_d = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \psi_d)$ et $\widehat{\boldsymbol{\beta}} = (\sum_{d=1}^{D} \hat{\gamma}_d \boldsymbol{x}_d \boldsymbol{x}_d{}')^{-1} \sum_{d=1}^{D} \hat{\gamma}_d \boldsymbol{x}_d \hat{\delta}_d^{DIR}$. Dans ce papier, dans un souci de concision, nous appellerons estimateur FH l'estimateur EBLUP fondé sur le modèle FH donné en (24).

Si les paramètres du modèle $\boldsymbol{\beta}$ et σ_u^2 sont connus, l'EQM de l'estimateur BLUP, $\tilde{\delta}_d^{FH}$, fondée sur le modèle (21) s'écrit

$$\mathrm{MSE}(\tilde{\delta}_d^{FH}) = \gamma_d \psi_d \leq \psi_d = \mathrm{var}_\pi(\hat{\delta}_d^{DIR}|\delta_d).$$

Par conséquent, étant donnée la vraie valeur de l'indicateur δ_d, si σ_u^2 et $\boldsymbol{\beta}$ sont connus, l'estimateur BLUP sous le modèle FH, $\tilde{\delta}_d^{FH}$, ne peut pas être moins efficace que l'estimateur direct. En pratique, σ_u^2 et $\boldsymbol{\beta}$ sont estimés et l'erreur due à l'estimation de ces deux paramètres s'ajoute à l'EQM de l'estimateur FH. Cependant, ces deux termes ajoutés à l'EQM tendent vers zéro lorsque le nombre de zones D tend vers l'infini. Par conséquent, pour un nombre suffisant de zones D, l'estimateur FH est toujours susceptible d'améliorer l'estimateur direct en termes d'EQM. C'est pourquoi ces estimateurs ont tendance à fournir une amélioration dans la plupart des domaines, tant que le nombre de domaines est suffisant. Cependant, les améliorations de l'efficacité seront faibles si le nombre de domaines n'est pas suffisamment important.

[4] Empirical BLUP.

Les modèles au niveau de l'unité élémentaire, fondés sur la taille totale de l'échantillon n, peuvent être beaucoup plus efficaces que les modèles au niveau du domaine, tant qu'il existe des variables auxiliaires au niveau individuel qui sont suffisamment informatives sur la variable expliquée. Cependant, un avantage de l'estimateur FH mis en évidence dans (24) est qu'il utilise les poids du plan d'échantillonnage par le biais de l'estimateur direct et qu'il est convergent sous le plan lorsque la taille de l'échantillon du domaine n_d augmente, alors que le poids de l'estimateur direct est $\gamma_d > 0$. En outre, son biais absolu sous le plan s'exprime comme suit

$$(1-\gamma_d)|\delta_d - \boldsymbol{x}_d{}'\boldsymbol{\beta}| \leq |\delta_d - \boldsymbol{x}_d{}'\boldsymbol{\beta}|,$$

ainsi, il sera moins biaisé que l'estimateur synthétique par régression fondé sur le même vecteur de coefficients $\boldsymbol{\beta}$ puisque $\gamma_d > 0$.

Pour un domaine non échantillonné, c'est-à-dire avec une taille d'échantillon $n_d = 0$, la variance de l'estimateur direct ψ_d tendrait vers l'infini et γ_d vers zéro. En supposant que la valeur limite est $\gamma_d = 0$, on obtient l'estimateur synthétique par régression

$$\hat{\delta}_d^{FH} = \boldsymbol{x}_d{}'\widehat{\boldsymbol{\beta}}.$$

Sous l'hypothèse de normalité de u_d et e_d, Prasad et Rao (1990) ont obtenu une approximation du second ordre (i.e., avec une erreur $o(D^{-1})$ quand le nombre de domaines D est grand) pour l'EQM de l'estimateur FH, qui s'écrit :

$$\mathrm{MSE}(\hat{\delta}_d^{FH}) = g_{d1}(\sigma_u^2) + g_{d2}(\sigma_u^2) + g_{d3}(\sigma_u^2),$$

où

$$\begin{aligned} g_{1d}(\sigma_u^2) &= \gamma_d \psi_d, \\ g_{2d}(\sigma_u^2) &= (1-\gamma_d)^2 \boldsymbol{x}_d{}' \left(\sum_{d=1}^{D} \gamma_d \, \boldsymbol{x}_d \boldsymbol{x}_d{}' \right)^{-1} \boldsymbol{x}_d, \\ g_{3d}(\sigma_u^2) &= (1-\gamma_d)^2 (\sigma_u^2 + \psi_d^2)^{-1} \overline{\mathrm{var}}(\hat{\sigma}_u^2). \end{aligned}$$

Ici, $\overline{\mathrm{var}}(\hat{\sigma}_u^2)$ est la variance asymptotique de l'estimateur $\hat{\sigma}_u^2$ de σ_u^2, qui dépend de la méthode d'estimation utilisée, $g_{1d}(\sigma_u^2)$ est l'erreur due à la prédiction de l'effet aléatoire du domaine u_d, de l'ordre de $O(1)$ lorsque D croît (c'est-à-dire ne tend pas vers zéro), $g_{2d}(\sigma_u^2)$ est l'erreur due à l'estimation du vecteur des coefficients de régression $\boldsymbol{\beta}$ et $g_{3d}(\sigma_u^2)$ est l'erreur due à l'estimation de la variance σ_u^2, où les deux derniers termes tendent vers zéro lorsque D croît de l'ordre $O(D^{-1})$; c'est-à-dire à la même vitesse que D^{-1}. Cela signifie que $g_{2d}(\sigma_u^2)$ et $g_{3d}(\sigma_u^2)$ disparaissent pour un D suffisamment grand, tandis que $g_{1d}(\sigma_u^2)$ ne disparaît pas, mais pour un D modéré, les trois termes doivent être pris en compte pour éviter une sous-estimation de l'EQM.

Si $\hat{\sigma}_u^2$ est l'estimateur REML, la variance asymptotique s'obtient comme inverse de l'information de Fisher $\mathcal{I}(\sigma_u^2)$, et s'exprime sous la forme :

$$\overline{\mathrm{var}}(\hat{\sigma}_u^2) = \mathcal{I}^{-1}(\sigma_u^2) = 2\left\{ \sum_{d=1}^{D} (\sigma_u^2 + \psi_d)^{-2} \right\}^{-1}. \qquad (25)$$

Dans ce cas, $g_{d2}(\hat{\sigma}_u^2)$ et $g_{d3}(\hat{\sigma}_u^2)$ sont les estimateurs respectifs de $g_{2d}(\sigma_u^2)$ et $g_{3d}(\sigma_u^2)$, sans biais au second ordre. Cela signifie que leur biais est $o(D^{-1})$, i.e., tend vers 0 plus vite que D^{-1} quand D croît. Cependant, $g_{d1}(\hat{\sigma}_u^2)$ a un biais non négligeable en tant qu'estimateur de $g_{d1}(\sigma_u^2)$ qui tend à être égal à $-g_{3d}(\sigma_u^2) + o(D^{-1})$. De ce fait, pour corriger le biais de $g_{1d}(\hat{\sigma}_u^2)$, on doit additionner deux fois $g_{3d}(\hat{\sigma}_u^2)$. Ainsi, un estimateur de l'EQM de l'estimateur FH non biaisé au second ordre, désigné ici sous le nom d'estimateur de Prasad-Rao, a pour expression

$$\text{mse}_{PR}(\hat{\delta}_d^{FH}) = g_{d1}(\hat{\sigma}_u^2) + g_{d2}(\hat{\sigma}_u^2) + 2g_{d3}(\hat{\sigma}_u^2).$$

Si $\hat{\sigma}_u^2$ est l'estimateur ML, sa variance asymptotique est la même que pour l'estimateur REML, donnée en (25). Cependant, cet estimateur a un biais qui s'écrit :

$$b(\sigma_u^2) = -\{2\mathcal{I}(\sigma_u^2)\}^{-1}\text{trace}\left[\left\{\sum_{d=1}^{D}(\sigma_u^2 + \psi_d)^{-1}\boldsymbol{x}_d\boldsymbol{x}_d{}'\right\}^{-1}\sum_{d=1}^{D}(\sigma_u^2 + \psi_d)^{-2}\boldsymbol{x}_d\boldsymbol{x}_d{}'\right].$$

Dans ce cas, le biais de l'estimateur ML ajoute un terme au biais de $g_{d1}(\hat{\sigma}_u^2)$ en tant qu'estimateur de $g_{d1}(\sigma_u^2)$. Ce biais est égal à $b(\sigma_u^2)\nabla g_{1d}(\sigma_u^2) - g_{3d}(\sigma_u^2)$, où

$$\nabla g_{1d}(\sigma_u^2) = (1 - \gamma_d)^2.$$

Puisque $b(\hat{\sigma}_u^2)\nabla g_{1d}(\hat{\sigma}_u^2)$ est un estimateur non biaisé au second ordre de $b(\sigma_u^2)\nabla g_{1d}(\sigma_u^2)$, nous pouvons corriger le biais de $g_{1d}(\hat{\sigma}_u^2)$ en soustrayant ce terme. De cette façon, on obtient l'estimateur non biaisé au second ordre de l'EQM de l'estimateur FH, ayant l'expression suivante,

$$\text{mse}_{PR}(\hat{\delta}_d^{FH}) = g_{d1}(\hat{\sigma}_u^2) - b(\hat{\sigma}_u^2)\nabla g_{1d}(\hat{\sigma}_u^2) + g_{d2}(\hat{\sigma}_u^2) + 2g_{d3}(\hat{\sigma}_u^2). \quad (26)$$

Si $\hat{\sigma}_u^2$ est l'estimateur obtenu par la méthode FH fondée sur les moments, l'estimateur non biaisé au second ordre de l'EQM a la même forme que (26), mais le biais de l'estimateur FH de σ_u^2 et la variance asymptotique changent, et s'écrivent

$$\overline{\text{var}}(\hat{\sigma}_u^2) = 2D\left\{\sum_{d=1}^{D}(\sigma_u^2 + \psi_d)^{-1}\right\}^{-2}, \quad (27)$$

$$b(\sigma_u^2) = \frac{2[D\sum_{d=1}^{D}(\sigma_u^2 + \psi_d)^{-2} - \{\sum_{d=1}^{D}(\sigma_u^2 + \psi_d)^{-1}\}^2]}{\{\sum_{d=1}^{D}(\sigma_u^2 + \psi_d)^{-1}\}^3}.$$

Quand on estime l'indicateur FGT d'ordre α, $\delta_d = F_{\alpha d}$, en utilisant le modèle FH, les variables auxiliaires $\boldsymbol{x}_d$ doivent vérifier le modèle

$$F_{\alpha d} = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d, \quad d = 1, \dots, D, \quad (28)$$

et l'on suppose que l'estimateur direct $\hat{F}_{\alpha d}^{DIR}$ de $F_{\alpha d}$ satisfait

$$\hat{F}_{\alpha d}^{DIR} = F_{\alpha d} + e_d, \quad d = 1, \dots, D. \quad (29)$$

Le modèle linéaire mixte obtenu en combinant (28) et (29) s'écrit

$$\hat{F}_{\alpha d}^{DIR} = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d + e_d, \quad d = 1, \dots, D. \quad (30)$$

L'ajustement de ce modèle conduit à un estimateur BLUP de $F_{\alpha d} = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d$ qui serait

$$\tilde{F}_{\alpha d}^{FH} = \boldsymbol{x}_d{}'\tilde{\boldsymbol{\beta}} + \tilde{u}_d, \quad (31)$$

où, dans ce cas, $\tilde{u}_d = \gamma_d(\hat{F}_{\alpha d}^{DIR} - \boldsymbol{x}_d{}'\tilde{\boldsymbol{\beta}})$ est l'estimateur BLUP de u_d et $\tilde{\boldsymbol{\beta}}$ est calculé comme suit :

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D}\gamma_d\,\boldsymbol{x}_d\boldsymbol{x}_d{}'\right)^{-1}\sum_{d=1}^{D}\gamma_d\,\boldsymbol{x}_d\hat{F}_{\alpha d}^{DIR}.$$

L'estimateur final FH de $F_{\alpha d}$ s'obtient en remplaçant simplement la variance σ_u^2 par un estimateur convergent $\hat{\sigma}_u^2$ dans l'expression de l'estimateur BLUP (31).

Les caractéristiques de l'estimateur FH peuvent être résumées comme suit:

Indicateurs cibles : paramètres généraux.

Données nécessaires :

- Données agrégées (par exemple, moyennes dans la population) des p variables auxiliaires considérées dans les domaines, $\boldsymbol{x}_d$, $d = 1, \dots, D$.

Avantages:

- Cet estimateur améliore généralement l'efficacité de l'estimateur direct.
- Le modèle de régression considéré intègre de l'hétérogénéité inexpliquée entre les domaines.
- Il s'agit d'un estimateur composite qui emprunte automatiquement des informations aux autres domaines (en donnant plus de poids à l'estimateur synthétique par la régression) lorsque cela est nécessaire (lorsque l'estimateur direct a une plus grande variance ou une plus petite taille d'échantillon). Il tend vers l'estimateur direct lorsque la taille du domaine augmente (lorsque ψ_d devient petit).
- Si, pour un domaine d, le poids donné à l'estimateur direct est strictement positif ($\gamma_d > 0$), les poids d'échantillonnage w_{di} sont utilisés au travers de l'estimateur direct $\hat{\delta}_d^{DIR}$; c'est-à-dire que le plan de sondage est pris en compte. Par conséquent, l'estimateur est convergent sous le plan de sondage (comme l'estimateur direct). Cela signifie qu'il sera moins affecté par les plans informatifs (plans avec des probabilités de sélection des individus dépendant de la variable d'intérêt), dans la mesure où les poids d'échantillonnage sont les vrais.
- Du fait de l'utilisation de données agrégées, l'estimateur FH n'est pas trop affecté par des valeurs aberrantes isolées (dans ce cas, des estimations directes atypiques pour un domaine).
- En n'utilisant que des informations auxiliaires agrégées, l'estimateur évite les problèmes de confidentialité des microdonnées obtenues à partir d'un recensement ou de fichiers administratifs.
- Pour des estimateurs linéaires directs, le théorème central limite s'applique pour les domaines ayant une taille d'échantillon suffisante. Ainsi, le modèle aura toujours une qualité d'ajustement minimale pour les domaines de taille d'échantillon suffisante.
- L'estimateur peut être mis en place dans des domaines non échantillonnés.
- L'estimateur de Prasad-Rao pour l'EQM est stable (ou efficace) et il est sans biais sous le plan quand il est moyenné sur de nombreux domaines.

Inconvénients :

- Les estimateurs sont fondés sur un modèle ; il est donc nécessaire d'analyser le modèle (par exemple, à l'aide des résidus). Il peut y avoir des problèmes dans le cas de paramètres non linéaires.
- Les variances d'échantillonnage des estimateurs directs ψ_d sont supposées connues, bien qu'en pratique il soit nécessaire de les estimer, ce qui conduit au même problème de l'absence de données dans un domaine. L'incorporation de l'erreur d'estimation de ces variances dans l'EQM de l'estimateur FH n'est pas automatique et souvent l'EQM estimée n'intègre pas cette erreur.
- Le nombre d'observations utilisées pour ajuster le modèle est le nombre de domaines échantillonnés, qui est généralement beaucoup plus petit que la taille totale de l'échantillon n utilisée pour ajuster les modèles au niveau individuel. Ainsi, les paramètres du modèle sont estimés avec une efficacité moindre et les améliorations de l'efficacité par rapport aux estimateurs directs seront plus faibles qu'avec les modèles au niveau individuel (cette

efficacité augmente avec le nombre de domaines). Dans nos applications, nous avons obtenu des gains très faibles par rapport à l'estimateur direct.

- Lors de l'estimation de plusieurs indicateurs qui dépendent d'une variable commune (par exemple, $F_{\alpha d}$ pour différentes valeurs de α), contrairement aux méthodes fondées sur des modèles au niveau des unités élémentaires, la modélisation et la recherche de variables auxiliaires utiles sont nécessaires pour chacun des indicateurs séparément.
- L'estimateur de l'EQM sous le modèle de Prasad-Rao est correct sous le modèle avec normalité de u_d et e_d, et il n'est pas sans biais sous le plan pour l'EQM sur un domaine particulier.
- Une fois que le modèle a été ajusté au niveau du domaine, les estimateurs $\hat{\delta}_d^{FH}$ ne peuvent pas être désagrégés en sous-domaines ou en sous-zones à l'intérieur des domaines, à moins de trouver un nouveau modèle adapté à ce nouveau niveau ou, sinon, d'ajuster un modèle à effets aléatoires multiniveaux.
- Les estimations doivent être réajustées pour vérifier une propriété d'additivité, afin que la somme des totaux estimés dans les domaines d'une région plus grande corresponde à l'estimateur direct pour cette zone.

Exemple 5. Estimateurs FH du taux de pauvreté avec R. En poursuivant avec les exemples précédents, nous montrons comment obtenir des estimateurs FH du taux de pauvreté avec R pour les provinces. Tout d'abord, pour vérifier si l'hypothèse de normalité du modèle est satisfaite, nous pouvons analyser graphiquement la distribution des estimateurs directs du taux de pauvreté au moyen de l'histogramme :

```
hist(povinc.dir,prob=TRUE,main="",xlab="HT estimators pov.
incidence")
```

La forme de cet histogramme (non reproduit par souci de concision) est quelque peu asymétrique mais n'est pas trop éloignée d'une densité normale, ce qui est prévisible puisque le Théorème central limite s'applique aux estimateurs directs des domaines.

Ensuite, nous chargeons les ensembles de données et les tailles de population des provinces, ventilés par groupes de nationalité, d'âge et de statut professionnel (certains ont déjà été chargés dans les exemples précédents) :

```
data(sizeprov)
data(sizeprovnat)
data(sizeprovage)
data(sizeprovedu)
data(sizeprovlab)
```

Nous utilisons ces tailles de population pour calculer les proportions d'individus dans chaque catégorie au sein de chaque province. Ce seront nos variables explicatives dans un modèle de Fay-Herriot :

```
Nd<-sizeprov[,3]
Ndnat<-as.matrix(sizeprovnat[,-c(1,2)])
Ndage<-as.matrix(sizeprovage[,-c(1,2)])
Ndedu<-as.matrix(sizeprovedu[,-c(1,2)])
Ndlab<-as.matrix(sizeprovlab[,-c(1,2)])
Pdnat<-Ndnat/Nd
Pdage<-Ndage/Nd
```

```
Pdedu<-Ndedu/Nd
Pdlab<-Ndlab/Nd

# Matrice du plan pour le modèle FH
X<-cbind(const=rep(1,D),nat1=Pdnat[,1],Pdage[,3:5],Pdedu[,c(1,3)],Pdlab[,c(2,3)])
```

Nous appelons la fonction qui calcule les estimateurs FH du taux de pauvreté pour les provinces, en utilisant les estimateurs directs HT obtenus dans l'Exemple 1 et leurs variances d'échantillonnage correspondantes :

```
povinc.FH.res<-eblupFH(povinc.dir~X-1,vardir=povinc.dir.res$SD^2)
povinc.FH<-povinc.FH.res$eblup
```

En utilisant les coefficients de régression estimés obtenus par l'ajustement du modèle de Fay-Herriot, nous pouvons également calculer des estimateurs de régression synthétiques fondés sur le modèle au niveau du domaine :

```
povinc.rsyn1<-X%*%povinc.FH.res$fit$estcoef[,1]
```

Bien que ces estimateurs soient fondés sur l'estimateur des coefficients de régression obtenu à partir de l'ajustement du modèle de Fay-Herriot et non du modèle synthétique, il s'agit également d'estimateurs synthétiques car ils ne tiennent pas compte de l'hétérogénéité entre les zones qui n'est pas expliquée par les variables auxiliaires considérées. De plus, les estimateurs des coefficients de régression obtenus dans les deux modèles, en utilisant les mêmes variables auxiliaires, sont asymptotiquement équivalents. Ainsi, pour un grand nombre de domaines, ils seront tous deux très similaires.

Comme les estimateurs FH sont des estimateurs composites entre les estimateurs de régression directs et les estimateurs synthétiques, nous calculons les poids donnés aux estimateurs directs dans l'estimateur composite et nous en montrons un résumé descriptif :

```
gammad<-povinc.FH.res$fit$refvar/(povinc.FH.res$fit$refvar+povinc.dir.res$SD^2)
summary(gammad)
```

Résultat :

```
   Min.   1st Qu.  Median   Mean     3rd Qu.  Max.
0.4537   0.7182   0.8108   0.7906   0.8977   0.9477
```

Nous constatons que, contrairement aux estimateurs SSC, dans ce cas, le poids accordé à l'estimateur direct n'est pas égal à un pour toutes les provinces, bien qu'il doive prendre des valeurs proches de un pour certaines provinces.

Nous comparons maintenant graphiquement les estimations FH avec les estimations directes HT et synthétiques RSYN1 pour chaque province. Les provinces sont classées sur l'axe de la plus petite à la plus grande taille d'échantillon, et nous indiquons leurs tailles d'échantillon sur l'axe des abscisses :

```
o<-order(nd)
k<-6
M<-max(povinc.dir,povinc.FH,povinc.rsyn1)
m<-min(povinc.dir,povinc.FH,povinc.rsyn1)
plot(1:D,povinc.dir[o],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="Estimator",
 xaxt="n")
points(1:D,povinc.dir[o],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:D,povinc.FH[o],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:D,povinc.rsyn1[o],type="b",col=3,lty=3,pch=3,lwd=2)
```

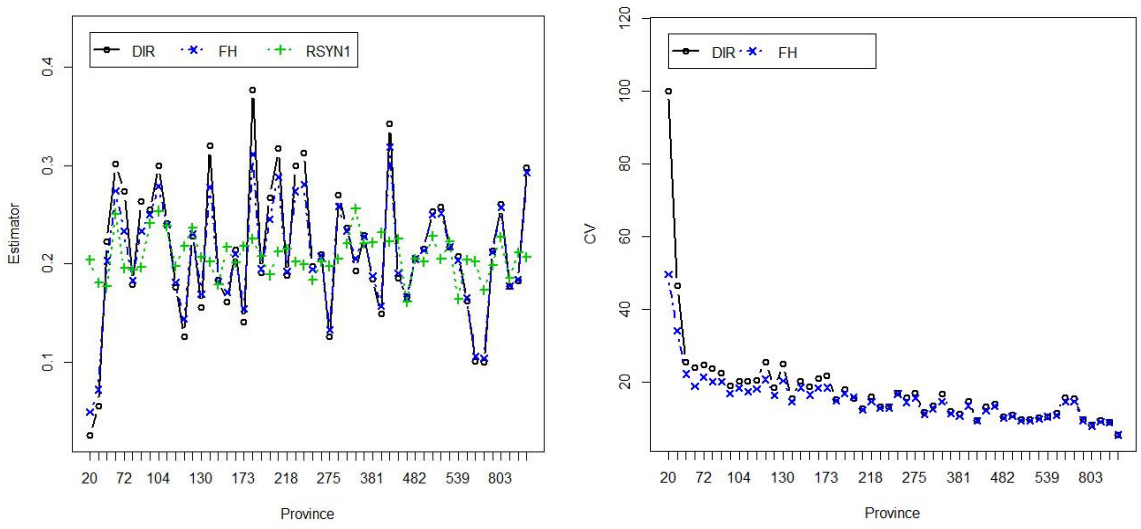
```
axis(1, at=1:D, labels=nd[o])
legend(1,M+(M-m)/k,legend=c("DIR","FH","RSYN1"),ncol=3,col=c(1,4,3),lwd=rep(2,3),
 lty=c(1,4,3),pch=c(1,4,3))
```

Enfin, nous estimons l'EQM des estimateurs FH en faisant appel à la fonction *mseFH()*, nous calculons les CV estimés et nous traçons les EQM en regard des variances des estimateurs directs :

```
povinc.FH.mse.res<-mseFH(povinc.dir~X-1,vardir=povinc.dir.res$SD^2)

povinc.FH.mse<-povinc.FH.mse.res$mse
povinc.FH.cv<-100*sqrt(povinc.FH.mse)/povinc.FH


M<-max(povinc.dir.var,povinc.FH.mse)
m<-min(povinc.dir.var,povinc.FH.mse)
plot(1:D,povinc.dir.cv[o],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="CV",xaxt="n")
points(1:D,povinc.dir.var[o],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:D,povinc.FH.mse[o],type="b",col=4,lty=4,pch=4,lwd=2)
axis(1, at=1:D, labels=nd[o])
legend(1,M+(M-m)/k,legend=c("DIR", "FH"),ncol=3,col=c(1,4),lwd=rep(2,2),lty=c(1,4),pch=c(1,4))
```

À nouveau, nous pouvons voir dans le graphique 6 (gauche) que les estimateurs synthétiques de régression prennent des valeurs très similaires pour toutes les provinces, contrairement aux estimateurs directs, qui varient davantage d'une province à l'autre. Les estimateurs FH sont proches des estimateurs directs, mais ils empruntent également des informations aux autres provinces par le biais des estimateurs synthétiques, en particulier pour les provinces dont la taille d'échantillon est plus petite (graphique de gauche).

Figure 6
Estimations FH, HT direct et RSYN1 du taux de pauvreté pour les provinces (à gauche), et EQM estimées à partir des estimateurs FH et HT direct (à droite)
(En proportions)



Source : D'après l'auteur.

Bien que dans cet exemple les variables auxiliaires considérées ne soient pas très efficientes, la figure 6 (à droite) indique que les estimateurs FH sont plus efficaces que les estimateurs directs.

Enfin, nous comparons les CVs estimés pour les estimateurs HT, GREG, et FH pour les 5 provinces ayant les plus petites tailles d'échantillon :

```
compardirFH<-data.frame(povinc.dir.cv,povinc.greg.cv,povinc.FH.cv)

selprov<-o[1:5]
compardirFH[selprov,]
```

Résultats :

```
   povinc.dir.CV povinc.greg.cv povinc.FH.cv
42      99.97815       94.72703     49.34572
5       46.35946       42.04802     33.74811
40      25.33449       21.77035     21.64444
34      23.80085       19.02477     18.27171
44      24.57017       16.86049     20.47468
```

Nous pouvons constater la réduction des CVs obtenue pour les estimateurs FH par rapport aux estimateurs directs HT. Ils sont également plus efficaces que les estimateurs GREG pour les quatre provinces ayant les tailles d'échantillon les plus petites, et ces améliorations sont significatives pour les deux provinces ayant les tailles d'échantillon les plus petites.

B. Estimateur EBLUP fondé sur le modèle avec erreurs emboîtées

Le modèle à erreurs emboîtées a été proposé par Battese, Harter et Fuller (1977) pour estimer la production de maïs et de soja au niveau d'un comté aux États-Unis. Ce modèle met en relation linéaire les valeurs d'une variable d'intérêt Y_{di} pour l'individu i dans le domaine d, avec les valeurs de p variables auxiliaires pour ce même individu, comme suit :

$$Y_{di} = \boldsymbol{x}_{di}{}'\boldsymbol{\beta} + u_d + e_{di}, \quad i = 1, \dots, N_d,\ d = 1, \dots, D, \tag{32}$$

où $\boldsymbol{\beta}$ est le vecteur des coefficients des variables auxiliaires, commun à tous les domaines, u_d est l'effet aléatoire du domaine et e_{di} est l'erreur au niveau individuel. Les effets aléatoires représentent l'hétérogénéité inexpliquée des valeurs des Y_{di} entre les domaines. Les effets aléatoires sont supposés indépendants des erreurs, avec $u_d \sim^{iid}(0, \sigma_u^2)$ et $e_{di} \sim^{ind}(0, \sigma_e^2 k_{di}^2)$, où les k_{di} sont des constantes connues traduisant une éventuelle hétéroscédasticité.

Notons que la moyenne sur le domaine d peut se décomposer en la somme des valeurs observées sur l'échantillon et celle sur les unités non échantillonnées, comme suit :

$$\bar{Y}_d = N_d^{-1}\left(\sum_{i \in s_d} Y_{di} + \sum_{i \in r_d} Y_{di}\right).$$

Il n'est pas nécessaire de prédire les valeurs observées dans l'échantillon puisqu'elles nous sont données. L'estimateur BLUP de $\bar{Y}_d$ sous le modèle avec erreurs emboîtées (32) s'obtient en ajustant simplement le modèle aux données issues de l'échantillon et en prédisant les valeurs des variables hors échantillon Y_{di} relatives au domaine d, i.e.

$$\tilde{\bar{Y}}_d^{BLUP} = N_d^{-1}\left(\sum_{i \in s_d} Y_{di} + \sum_{i \in r_d} \tilde{Y}_{di}^{BLUP}\right), \tag{33}$$

où, en prenant l'estimateur des moindres carrés pondérés de $\boldsymbol{\beta}$, soit $\tilde{\boldsymbol{\beta}}$ issu du modèle (32), les valeurs prédites sont

$$\tilde{Y}_{di}^{BLUP} = \boldsymbol{x}_{di}{}'\tilde{\boldsymbol{\beta}} + \tilde{u}_d,$$
$$\tilde{u}_d = \gamma_d(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}{}'\tilde{\boldsymbol{\beta}}), \gamma_d = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/a_{d\cdot}),$$

où $\bar{y}_{da} = a_{d\cdot}^{-1}\sum_{i\in s_d} a_{di}\, Y_{di}$ et $\bar{\boldsymbol{x}}_{da} = a_{d\cdot}^{-1}\sum_{i\in s_d} a_{di}\, \boldsymbol{x}_{di}$ sont les moyennes pondérées dans l'échantillon de la variable expliquée et des variables auxiliaires, respectivement, avec les poids $a_{di} = k_{di}^{-2}$, et où $a_{d\cdot} = \sum_{i\in s_d} a_{di}$. A nouveau $\tilde{u}_d$ est l'estimateur BLUP de u_d et les valeurs prédites $\tilde{Y}_{di}^{BLUP}$ sont les estimations BLUP des variables Y_{di}, $i \in r_d$, sous le modèle (32).

Nous construisons le vecteur des variables expliquées pour le domaine d, $\boldsymbol{y}_d = (Y_{d1}, \dots, Y_{dN_d})'$ et la matrice correspondante des variables auxiliaires, $\boldsymbol{X}_d = (\boldsymbol{x}_{d1}, \dots, \boldsymbol{x}_{dN_d})'$. Sous le modèle à erreurs emboîtées (32), $\boldsymbol{y}_d \sim^{ind} N(\boldsymbol{X}_d\boldsymbol{\beta}, \boldsymbol{V}_d)$, $d = 1, \dots, D$, où

$$\boldsymbol{V}_d = \sigma_u^2\mathbf{1}_{N_d}\mathbf{1}_{N_d}' + \sigma_e^2\boldsymbol{A}_d,$$

où $\boldsymbol{A}_d = \mathrm{diag}(k_{di}^2; i = 1, \dots, N_d)$. Décomposons maintenant le vecteur $\boldsymbol{y}_d$ relatif au domaine d en sous-vecteurs relatifs aux unités échantillonnées d'une part et à celles hors échantillon, d'autre part, comme suit : $\boldsymbol{y}_d = (\boldsymbol{y}_{ds}{}', \boldsymbol{y}_{dr}{}')'$, et, de même pour les matrices $\boldsymbol{X}_d$ and $\boldsymbol{V}_d$,

$$\boldsymbol{X}_d = \begin{pmatrix}\boldsymbol{X}_{ds}\\ \boldsymbol{X}_{dr}\end{pmatrix}, \quad \boldsymbol{V}_d = \begin{pmatrix}\boldsymbol{V}_{ds} & \boldsymbol{V}_{dsr}\\ \boldsymbol{V}_{drs} & \boldsymbol{V}_{dr}\end{pmatrix}.$$

Avec cette notation, l'estimateur des moindres carrés pondérés de $\boldsymbol{\beta}$ s'écrit :

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \boldsymbol{X}_{ds}\,\boldsymbol{V}_{ds}^{-1}\boldsymbol{X}_{ds}{}'\right)^{-1}\sum_{d=1}^{D} \boldsymbol{X}_{ds}\,\boldsymbol{V}_{ds}^{-1}\boldsymbol{y}_{ds}. \tag{34}$$

Pour des domaines où la fraction de l'échantillon est très faible, c'est-à-dire $n_d/N_d \approx 0$, l'estimateur BLUP de la moyenne $\bar{Y}_d$ peut s'écrire comme suit :

$$\tilde{\bar{Y}}_d^{BLUP} \approx \gamma_d\{\bar{y}_{da} + (\bar{\boldsymbol{X}}_d - \bar{\boldsymbol{x}}_{da})'\tilde{\boldsymbol{\beta}}\} + (1-\gamma_d)\bar{\boldsymbol{X}}_d{}'\tilde{\boldsymbol{\beta}}.$$

Comme $\gamma_d \in (0,1)$, l'estimateur BLUP est une moyenne pondérée de l'estimateur $\bar{y}_{da} + (\bar{\boldsymbol{X}}_d - \bar{\boldsymbol{x}}_{da})'\tilde{\boldsymbol{\beta}}$, connu sous le nom d'estimateur par régression sur l'enquête, et de l'estimateur synthétique par régression, $\bar{\boldsymbol{X}}_d{}'\tilde{\boldsymbol{\beta}}$. L'estimateur par régression sur l'enquête est obtenu en adaptant le même modèle (32) mais en considérant les effets du domaine u_d comme fixes plutôt qu'aléatoires. Notons aussi que cette moyenne pondérée est similaire à celle obtenue en utilisant l'estimateur FH donné en (24), mais où l'estimateur par régression sur l'enquête $\bar{y}_{da} + (\bar{\boldsymbol{X}}_d - \bar{\boldsymbol{x}}_{da})'\tilde{\boldsymbol{\beta}}$ joue le rôle d'estimateur direct. En fait, cet estimateur peut être considéré comme direct, puisque sa variance est $O(n_d^{-1})$; i.e., sa variance croît quand la taille d'échantillon du domaine n_d devient petite.

Pour interpréter cet estimateur, considérons, pour simplifier, un modèle homoscédastique ; i.e., avec $k_{di} = 1$ pour tout i et d. Dans ce cas, on a $\gamma_d = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/n_d)$. Pour un domaine avec une petite taille d'échantillon n_d, γ_d est proche de zéro et l'estimateur BLUP est proche de l'estimateur synthétique par la régression, qui utilise des informations provenant des autres domaines. Cependant, pour un domaine avec un échantillon de grande taille n_d, γ_d est proche de un et l'estimateur BLUP est proche de l'estimateur par régression sur l'enquête. De plus, γ_d dépend également de l'hétérogénéité entre les domaines, mesurée par σ_u^2.Si les domaines sont très hétérogènes (σ_u^2 est grand par rapport à σ_e^2/n_d), ou, de manière équivalente, si les variables auxiliaires considérées n'expliquent pas une grande partie de la variabilité, alors γ_d est proche de un et on donne plus de poids à l'estimateur par régression sur l'enquête, qui est similaire à un estimateur direct. Dans le cas contraire, si les zones sont homogènes ou, en d'autres termes, si les variables auxiliaires sont des prédicteurs forts, alors on donne plus de poids à l'estimateur synthétique obtenu par régression avec ces variables auxiliaires.

À nouveau, l'estimateur BLUP donné en (33) dépend des vraies valeurs des composantes de la variance du modèle (32), $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)'$. En remplaçant la vraie valeur $\boldsymbol{\theta}$ par un estimateur convergent $\widehat{\boldsymbol{\theta}} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ dans l'expression de l'estimateur BLUP (33), nous obtenons l'estimateur EBLUP, qui s'écrit :

$$\hat{\bar{Y}}_d^{EBLUP} = N_d^{-1}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} \hat{Y}_{di}^{EBLUP}\right), \tag{35}$$

où, si $\widehat{\boldsymbol{\beta}}$ est le résultat de la substitution à $\boldsymbol{\theta}$ de l'estimateur $\widehat{\boldsymbol{\theta}}$ dans l'expression de $\widetilde{\boldsymbol{\beta}}$ donnée en (34), les valeurs prédites deviennent

$$\hat{Y}_{di}^{EBLUP} = \boldsymbol{x}_{di}'\widehat{\boldsymbol{\beta}} + \hat{u}_d,$$
$$\hat{u}_d = \hat{\gamma}_d(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}'\widehat{\boldsymbol{\beta}}), \hat{\gamma}_d = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_e^2/a_{d\cdot}).$$

L'estimateur BLUP est sans biais sous le modèle (32) et il est optimal, en ce sens qu'il minimise l'EQM, entre les estimateurs linéaires sur l'échantillon et sans biais. En remplaçant $\boldsymbol{\theta}$ par l'estimateur $\widehat{\boldsymbol{\theta}}$, l'estimateur EBLUP reste non biaisé sous le modèle (32), sous certaines conditions relatives à l'estimateur $\widehat{\boldsymbol{\theta}}$. Les méthodes d'estimation habituelles, à savoir ML, REML et la méthode de Henderson III, satisfont ces conditions. Cependant, ni l'estimateur BLUP ni le EBLUP ne sont sans biais sous le plan d'échantillonnage. En fait, ils ne tiennent pas compte du plan d'échantillonnage et sont donc normalement conçus pour un échantillonnage aléatoire simple (SRS[5]). Dans tous les cas, les estimateurs EBLUP fournissent une nette amélioration de l'efficacité par rapport aux estimateurs directs et même par rapport aux estimateurs FH, puisqu'ils utilisent des informations beaucoup plus détaillées et de manière plus efficace (sans réduire de moitié les données). Dans le cadre de plans d'échantillonnage à probabilités inégales, ils peuvent présenter un biais non négligeable sous le plan. You et Rao (2002) ont proposé une variante appelée pseudo-EBLUP qui prend en compte les poids d'échantillonnage et qui est convergente sous le plan de sondage lorsque la taille du domaine n_d augmente.

Pour un domaine non échantillonné, i.e., avec une taille d'échantillon $n_d = 0$, en supposant $\gamma_d = 0$, on obtient l'estimateur synthétique par régression $\bar{\boldsymbol{X}}_d'\widehat{\boldsymbol{\beta}}$.

Sous un plan de sondage aléatoire simple et en supposant $k_{di} = 1$, pour tout i et d, étant donné que l'estimateur par régression sur l'enquête est approximativement sans biais sous le plan, le biais sous le modèle BLUP quand $n_d/N_d \approx 0$ est $-(1-\gamma_d)(\bar{Y}_d - \bar{\boldsymbol{X}}_d'\boldsymbol{\beta})$. De ce fait, le biais relatif en valeur absolue (RAB[6]) sous le plan est égal à

$$(1-\gamma_d)\left|\frac{\bar{Y}_d - \bar{\boldsymbol{X}}_d'\boldsymbol{\beta}}{\bar{Y}_d}\right| \leq \left|\frac{\bar{Y}_d - \bar{\boldsymbol{X}}_d'\boldsymbol{\beta}}{\bar{Y}_d}\right|,$$

i.e., il est plus petit que la valeur absolue du biais relatif sous le plan de l'estimateur synthétique par régression $\bar{\boldsymbol{X}}_d'\boldsymbol{\beta}$ pour le même vecteur de coefficients $\boldsymbol{\beta}$, $|(\bar{Y}_d - \bar{\boldsymbol{X}}_d'\boldsymbol{\beta})/\bar{Y}_d|$, tant que $\gamma_d > 0$. Si nous fixons une borne supérieure B pour le biais relatif en valeur absolue (par exemple, $B = 0.20$ ou $B = 0.10$), si cette limite B est dépassée pour certains domaines, nous pouvons remplacer le biais relatif absolu de l'estimateur synthétique par une quantité constante pour chaque domaine, telle que le maximum, i.e., nous posons

$$M = \max_{1\leq d\leq D}\left|\frac{\bar{Y}_d - \bar{\boldsymbol{X}}_d'\boldsymbol{\beta}}{\bar{Y}_d}\right|.$$

[5] Simple random sampling.
[6] Relative absolute bias.

La quantité $(1-\gamma_d)M$ décroît régulièrement avec la taille d'échantillon du domaine n_d, par l'intermédiaire de γ_d. On peut trouver la taille d'échantillon n_d^* à partir de laquelle $(1-\gamma_d)M$ dépasse B. Si $M > B$ (sinon le RAB ne dépasse pas B, pour aucune province), la taille d'échantillon résultante est

$$n_d^* = \frac{\sigma_e^2}{\sigma_u^2}\left(\frac{M}{B} - 1\right).$$

Ainsi, pour des domaines dont la taille d'échantillon est $n_d < n_d^*$, le biais relatif absolu pourrait aller au-delà de la borne supérieure B et l'on pourrait décider de ne pas fournir d'estimations pour ces domaines.

Cependant, n_d^* dépend de certaines quantités inconnues. Par conséquent, en pratique, nous estimons ces quantités inconnues et obtenons une valeur estimée de n_d^*. Un estimateur serait

$$\hat{n}_d^* = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_u^2}\left(\frac{\widehat{M}}{B} - 1\right),$$

où

$$\widehat{M} = \max_{1\le d\le D}\left|\frac{\hat{\bar{Y}}_d^{EBLUP} - \bar{\boldsymbol{X}}_d{}'\widehat{\boldsymbol{\beta}}}{\hat{\bar{Y}}_d^{EBLUP}}\right|,$$

en supposant que $\widehat{M} > B$.

L'EQM de l'estimateur EBLUP $\hat{\bar{Y}}_d^{EBLUP}$ de $\bar{Y}_d$, ainsi qu'un estimateur du second ordre de cette EQM, peuvent être approximés en utilisant une formule analytique appropriée du second ordre pour D, quasiment de la même manière que la formule de Prasad-Rao décrite dans l'introduction pour la partie consacrée à l'estimateur FH. Une autre option qui ne nécessite pas un grand nombre de domaines D, bien que plus coûteuse en termes de calcul, consiste à se tourner vers les procédures de Bootstrap.Nous donnons ici un aperçu d'une procédure de Bootstrap paramétrique pour les populations finies proposée par González-Manteiga et al. (2008), appliquée ici pour l'estimation des moyennes des zones, $\bar{Y}_d$. La procédure de Bootstrap est la suivante :

- Appliquer le modèle (32) aux données de l'échantillon $\boldsymbol{y}_s = (\boldsymbol{y}_{1s}{}', \dots, \boldsymbol{y}_{Ds}{}')'$ et obtenir les estimateurs des paramètres du modèle $\widehat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ et $\hat{\sigma}_e^2$.
- Générer les effets alaétoires sur les domaines comme suit : $u_d^{*(b)} \sim^{iid} N(0, \hat{\sigma}_u^2)$, $d = 1, \dots, D$.
- Indépendamment des effets aléatoires dans les domaines $u_d^{*(b)}$, générer des erreurs bootstrap pour les unités échantillonnées dans le domaine, $e_{di}^{*(b)} \sim^{iid} N(0, \hat{\sigma}_e^2)$, $i \in s_d$. Générer aussi les valeurs moyennes des erreurs dans la population pour chaque domaine, $\bar{E}_d^{*(b)} \sim^{iid} N(0, \hat{\sigma}_e^2/N_d)$, $d = 1, \dots, D$.
- Calculer les valeurs moyennes bootstrap pour les domaines,
$$\bar{Y}_d^{*(b)} = \bar{\boldsymbol{X}}_d{}'\widehat{\boldsymbol{\beta}} + u_d^{*(b)} + \bar{E}_d^{*(b)}, \quad d = 1, \dots, D.$$
Notons que le calcul de la moyenne $\bar{Y}_d^{*(b)}$ ne nécessite pas de connaître les valeurs individuelles $\boldsymbol{x}_{di}$, pour chaque unité hors échantillon du domaine $i \in r_d$.
- En utilisant les vecteurs des valeurs des variables auxiliaires pour les unités échantillonnées $\boldsymbol{x}_{di}$, $i \in s_d$, générer les variables expliquées pour les unités échantillonnées fondées sur le modèle
$$Y_{di}^{*(b)} = \boldsymbol{x}_{di}{}'\widehat{\boldsymbol{\beta}} + u_d^{*(b)} + e_{di}^{*(b)}, \quad i \in s_d, \quad d = 1, \dots, D. \qquad (36)$$
- Pour l'échantillon de départ $s = s_1 \cup \cdots \cup s_D$, soit $\boldsymbol{y}_s^{*(b)} = ((\boldsymbol{y}_{1s}^{*(b)})', \dots, (\boldsymbol{y}_{Ds}^{*(b)})')'$ le vecteur bootstrap des valeurs dans l'échantillon. Appliquer le modèle (32) aux données bootstrap $\boldsymbol{y}_s^{*(b)}$ et calculer les estimateurs EBLUP bootstrap s $\hat{\bar{Y}}_d^{EBLUP*(b)}$, $d = 1, \dots, D$.

Répéter les étapes 2) - 6) pour $b = 1, \dots, B$. On obtient ainsi les vraies valeurs des moyennes $\bar{Y}_d^{*(b)}$ et les estimateurs EBLUP correspondants $\hat{\bar{Y}}_d^{EBLUP*(b)}$ pour la réplication Bootstrap b. Les estimateurs naïfs de l'EQM des estimateurs EBLUP $\hat{\bar{Y}}_d^{EBLUP}$, obtenus au moyen de la méthode de Boostrap paramétrique sont

$$mse_B\left(\hat{\bar{Y}}_d^{EBLUP}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\bar{Y}}_d^{EBLUP*(b)} - \bar{Y}_d^{*(b)}\right)^2, \quad d = 1, \dots, D. \tag{37}$$

L'estimateur bootstrap (37) est sans biais au premier ordre plutôt qu'au second, i.e., son biais ne décroît pas plus vite que D^{-1} quand le nombre de domaines D augmente. Il y a diverses corrections de biais dans la littérature mais elles produisent des estimateurs qui peuvent prendre des valeurs négatives non adaptées ou bien elles sont strictement positives mais elles présentent un biais de second ordre. De plus, ces corrections augmentent la variance de l'estimateur de l'EQM. Ainsi, l'estimateur bootstrap naïf qui n'effectue pas de correction de biais constitue-t-il un choix acceptable parmi les estimateurs non analytiques.

Résumé des caractéristiques de la méthode EBLUP fondée sur le modèle avec erreurs emboîtées.

Indicateurs cibles : valeurs moyennes/totaux de la variable d'intérêt.

Données nécessaires :

- Microdonnées pour les p variables auxiliaires considérées, provenant de la même enquête que celle où la variable d'intérêt est observée.
- Domaines d'intérêt obtenus à partir de la même enquête que celle où la variable d'intérêt est observée.
- Valeurs moyennes dans la population des p variables auxiliaires considérées pour les différents domaines, $\bar{\boldsymbol{X}}_d$, $d = 1, \dots, D$.

Avantages :

- Le nombre d'observations utilisées pour ajuster le modèle est exactement la taille totale de l'échantillon n, beaucoup plus grande que le nombre d'observations (égal au nombre de domaines) dans les modèles FH. Ainsi, les paramètres du modèle sont estimés plus efficacement et l'amélioration de l'efficacité par rapport aux estimateurs directs sera plus importante qu'avec les modèles FH.
- Le modèle de régression considéré incorpore de l'hétérogénéité inexpliquée entre les domaines.
- Il s'agit d'un estimateur composite, qui utilise automatiquement des informations provenant des zones restantes (en donnant plus de poids à l'estimateur synthétique par régression) lorsque cela est nécessaire (lorsque la taille de l'échantillon est faible). Il tend vers l'estimateur synthétique par la régression sur l'enquête lorsque la taille du domaine augmente.
- Contrairement au modèle FH, il n'est pas nécessaire de connaître la variance.
- L'estimateur de l'EQM sous le modèle est un estimateur stable de l'EQM sous le plan de sondage et il est sans biais sous le plan lorsqu'il est moyenné sur plusieurs domaines.
- Les estimations peuvent être désagrégées pour tout sous-domaine ou sous-zone que l'on souhaite au sein des domaines, et même au niveau individuel.

- Il peut se calculer pour des domaines non échantillonnés.

Inconvénients :

- Les estimateurs sont fondés sur un modèle ; il est donc nécessaire d'analyser le modèle (par exemple, au moyen des résidus).
- Ils ne prennent pas en compte le plan d'échantillonnage. Par conséquent, ils ne sont pas sans biais sous le plan de sondage et ils conviennent mieux à l'échantillonnage aléatoire simple. Ils seront affectés par des plans d'échantillonnage informatifs.
- Ils sont affectés par des observations isolées et aberrantes ou par l'absence de normalité.
- Les microdonnées sont généralement obtenues à partir d'un recensement ou de fichiers administratifs, et il existe souvent des problèmes de confidentialité qui limitent l'utilisation de ce type de données.
- L'estimateur de l'EQM sous le modèle de Prasad-Rao convient sous le modèle avec normalité et il n'est pas sans biais sous le modèle pour l'EQM pour un domaine particulier.
- Ils nécessitent un réajustement pour vérifier la propriété d'additivité, de façon à ce que la somme des totaux estimés dans les domaines d'une plus grande région corresponde à l'estimateur direct pour cette zone.

Exemple 6. Estimateurs EBLUP fondés sur le modèle avec erreurs emboîtées pour le taux de pauvreté, avec R. En poursuivant les exemples précédents, nous montrons comment obtenir avec R les estimateurs EBLUP du taux de pauvreté fondés sur un modèle avec erreurs emboîtées. Dans un jeu de données prédéfini dans R, les valeurs des variables auxiliaires hors échantillon sont disponibles pour les cinq provinces ayant les plus petites tailles d'échantillon. En utilisant ces données et l'échantillon, nous pouvons calculer les moyennes dans la population de ces variables pour ces provinces, mais nous ne disposons pas des moyennes réelles pour les autres provinces. Par conséquent, nous ne pouvons mettre en évidence les estimateurs EBLUP que pour ces provinces, même si le modèle est ajusté à l'échantillon total.

Tout d'abord, nous chargeons le jeu de données contenant les valeurs des variables auxiliaires hors échantillon pour les provinces sélectionnées et nous calculons les moyennes dans la population de ces variables dans les provinces. Pour ce faire, nous utilisons les valeurs dans l'échantillon (jeu de données incomedata) et les valeurs hors échantillon (Xoutsamp). De plus, nous incluons les codes des provinces dans la première colonne de la matrice des valeurs moyennes :

```
data(Xoutsamp)

I<-length(selprov)                  # Nombre de provinces sélectionnées
p<-dim(Xoutsamp)[2]-1               # Nombre de variables auxiliaires

auxvar<-names(Xoutsamp)[-1]         # Noms des variables auxiliaires dans Xoutsamp
meanXpop<-matrix(0,nr=I,nc=p)       # Matrice des moyennes des variables auxiliaires
Ni<-numeric(I)                      # Taille de la population des provinces

for (i in 1:I){                     # Boucle pour les provinces sélectionnées
  d<-selprov[i]
  Xsd<-incomedata[prov==d,auxvar]           # Valeurs dans l'échantillon des variables auxiliaires
sampling values
  Xrd<-Xoutsamp[Xoutsamp$domain==d,-1]      # Valeurs hors échantillon
  Ni[i]<-dim(Xrd)[1]+dim(Xsd)[1]# Taille de la population de la prov.
```

```
 for (k in 1:p){
  meanXpop[i,k]<-(sum(Xrd[,k])+sum(Xsd[,k]))/Ni[i]
 }
}
Xmean<-data.frame(selprov,meanXpop)
```

Nous faisons maintenant appel à la fonction qui calcule les estimateurs EBLUP du taux de pauvreté pour les provinces sélectionnées, sur la base du modèle à erreurs emboîtées adapté aux données de l'échantillon (pour toutes les provinces). Nous sauvegardons les estimations obtenues dans un vecteur :

```
povinc.BHF.res<-eblupBHF(poor ~ age2+age3+age4+age5+nat1+educ1+educ3+labor1+labor2,
 dom=prov,selectdom=selprov,meanxpop=Xmean,popnsize=sizeprov[,-1])

povinc.BHF<-numeric(D)
povinc.BHF[selprov]<-povinc.BHF.res$eblup$eblup
```

Nous vérifions les résultats de l'ajustement du modèle avec erreurs emboîtées et calculons l'estimateur synthétique par la régression fondé sur le modèle au niveau individuel :

```
betaest<-povinc.BHF.res$fit$fixed      # Coefficients de la régression
upred<-povinc.BHF.res$fit$random   # Prédiction des effets des provinces
sigmae2est<-povinc.BHF.res$fit$errorvar  # Variance estimée de l'erreur
sigmau2est<-povinc.BHF.res$fit$refvar   # Variance estimée des effets des provinces

povinc.rsyn2<-numeric(D)
povinc.rsyn2[selprov]<-cbind(1,meanXpop)%*%betaest
```

Nous analysons le poids que l'estimateur EBLUP donne à l'estimateur par régression sur l'enquête:

```
gammad.BHF<-sigmau2est/(sigmau2est+sigmae2est/nd)
summary(gammad.BHF)
```

Résultat :

```
   Min.   1st Qu.   Median    Mean      3rd Qu.   Max.
 0.3458   0.7743    0.8606    0.8352    0.9276    0.9741
```

Plus le résultat de gammad.BHF est proche de zéro pour un domaine, plus d'information est tirée de l'estimateur synthétique par régression au niveau individuel. Dans ce cas, il y a une province pour laquelle beaucoup d'informations sont extraites de cet estimateur, étant donné que la valeur minimale de gammad.BHF est relativement faible.

Nous calculons maintenant les estimateurs de l'EQM des estimateurs EBLUP en utilisant le bootstrap paramétrique décrit ci-dessus. Pour ce faire, nous faisons appel à la fonction pbmseBHF() en utilisant B=200 réplications bootstrap. Cette fonction fournit également les estimations EBLUP et les résultats de l'ajustement exactement de la même manière que la fonction eblupBHF().

```
povinc.mse.res<-pbmseBHF(poor~age3+age4+age5+nat1+educ1+educ3+labor1+labor2,
dom=prov,selectdom=selprov,meanxpop=Xmean,popnsize=sizeprov[,-1],B=200)
```

Enfin, nous comparons les estimateurs EBLUP fondés sur le modèle avec erreurs emboîtées avec les estimateurs directs HT et FH en traçant les estimations ponctuelles obtenues et leurs EQM estimées pour les cinq provinces sélectionnées :

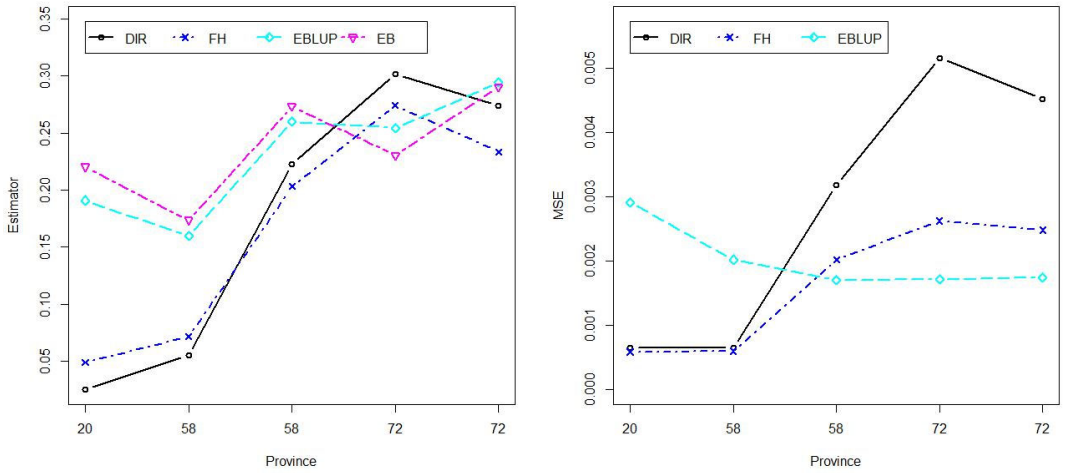
Ci-dessous le code R utilisé pour obtenir la figure suivante :

```
M<-max(povinc.dir[selprov],povinc.FH[selprov],povinc.BHF[selprov])
m<-min(povinc.dir[selprov],povinc.FH[selprov],povinc.BHF[selprov])
plot(1:5,povinc.dir[selprov],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="Estimator",
 xaxt="n")
points(1:5,povinc.dir[selprov],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:5,povinc.FH[selprov],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:5,povinc.BHF[selprov],type="b",col=5,lty=5,pch=5,lwd=2)
axis(1, at=1:5, labels=nd[selprov])
legend(1,M+(M-m)/k,legend=c("DIR","FH","EBLUP"),ncol=3,col=c(1,4,5),lwd=rep(2,3),
 lty=c(1,4,5),pch=c(1,4,5))

M<-max(povinc.dir.var[selprov],povinc.FH.mse[selprov],povinc.BHF.mse[selprov])
m<-min(povinc.dir.var[selprov],povinc.FH.mse[selprov],povinc.BHF.mse[selprov])
plot(1:5,povinc.dir.cv[selprov],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="CV",
 xaxt="n")
points(1:5,povinc.dir.var[selprov],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:5,povinc.FH.mse[selprov],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:5,povinc.BHF.mse[selprov],type="b",col=5,lty=5,pch=5,lwd=2)
axis(1, at=1:5, labels=nd[selprov])
legend(1,M+(M-m)/k,legend=c("DIR", "FH", "EBLUP"),ncol=3,col=c(1,4,5),lwd=rep(2,3),
 lty=c(1,4,5),pch=c(1,4,5))
```

D'après la figure 7 (à gauche), nous pouvons voir que, pour les cinq provinces ayant la plus petite taille d'échantillon, les estimateurs FH prennent des valeurs similaires aux estimateurs directs mais sont légèrement plus stables pour les 5 provinces sélectionnées que les estimateurs directs. Les EBLUP sont clairement plus stables pour les 5 provinces sélectionnées que les estimateurs directs et FH. De plus, comme nous pouvons l'observer dans la Figure 7 (à droite), les EQM estimées des estimateurs FH sont plus petites pour les provinces de gauche, car elles utilisent davantage d'informations provenant des autres provinces, puisque le modèle avec erreurs emboîtées est ajusté en prenant en compte tous les individus de l'échantillon (à partir de D=52 provinces).

D'autre part, ces EQM augmentent progressivement lorsque la taille de l'échantillon diminue, ce qui est logique. En revanche, les EQM estimées des estimateurs directs et FH prennent des valeurs extrêmement faibles pour les provinces dont la taille d'échantillon est plus petite (ce qui n'est pas très crédible). Dans le cas des estimateurs directs, leurs variances sont estimées avec les données restreintes à chaque province et, par conséquent, ces variances estimées (comme les EQM) ne sont pas fiables. Les estimateurs BLUP basés sur le modèle FH avec des paramètres connus ont une EQM qui ne peut pas dépasser la variance des estimateurs directs ; si ces variances sont incorrectement estimées, alors l'EQM de l'estimateur FH est également incorrecte dans ce cas.

Figure 7
Estimations EBLUP fondées sur le modèle avec erreurs emboîtées du taux de pauvreté pour les provinces ainsi qu'estimations directes HT et FH (à gauche), et EQM estimées pour les trois estimateurs (à droite)
(En proportions)



Source : D'après l'auteur.

C. Méthode ELL

La méthode d'Elbers, Lanjouw et Lanjouw (2003), que nous appellerons la méthode ELL, est la méthode traditionnellement utilisée par la Banque Mondiale pour construire des cartes de pauvreté ou d'inégalité. Cette méthode a été la première à apparaître dans la littérature pour permettre d'estimer des indicateurs plus complexes que les moyennes ou les totaux, pour autant qu'ils soient fonction d'une variable qui mesure le pouvoir d'achat individuel (généralement le revenu net disponible ou les dépenses). Cette méthode fait l'hypothèse d'un modèle à erreurs emboîtées (32) sur la variable transformée de façon logarithmique, dans lequel les effets aléatoires concernent les unités de premier degré du plan d'échantillonnage (grappes) plutôt que les domaines d'intérêt. Cependant, pour faciliter la comparaison avec les autres méthodes présentées dans ce document, tout en simplifiant la notation, nous considérerons que les unités de premier degré sont identiques aux domaines. Dans ce cas, si E_{di} est la variable qui mesure la pouvoir d'achat de l'individu i dans le domaine d, en supposant $Y_{di} = \log(E_{di} + c)$, où $c > 0$ est une constante, alors le modèle ELL est :

$$Y_{di} = \boldsymbol{x}_{di}'\boldsymbol{\beta} + u_d + e_{di}, \quad i = 1, \dots, N_d,\ d = 1, \dots, D, \tag{38}$$

où $u_d \sim^{iid}(0, \sigma_u^2)$ et $e_{di} \sim^{ind}(0, \sigma_e^2 k_{di}^2)$, u_d et e_{di} étant indépendants, et les k_{di} sont des constantes connues traduisant une éventuelle hétéroscédasticité.

L'estimateur ELL d'un paramètre général $\delta_d = \delta_d(\boldsymbol{y}_d)$ sous le modèle (38) est obtenu au moyen d'une procédure Bootstrap. Cette procédure Bootstrap fournit une approximation numérique de l'estimateur ELL théorique, qui s'exprime comme l'espérance totale $\hat{\delta}_d^{ELL} = E[\delta_d]$, contrairement au prédicteur EB considéré au chapitre V.B, qui implique des conditions sur l'échantillon $\boldsymbol{y}_s$. La même procédure Bootstrap est utilisée pour obtenir une estimation de l'EQM de l'estimateur ELL.

La procédure Bootstrap fonctionne comme suit. Tout d'abord, les résidus du modèle (38) estimé sur les données sont utilisés pour générer des effets aléatoires u_d^* pour chaque domaine $d = 1, \dots, D$, et des erreurs e_{di}^*, pour chaque individu $i = 1, \dots, N_d$, $d = 1, \dots, D$. À partir de ceux-ci et de l'estimateur $\widehat{\boldsymbol{\beta}}$ du paramètre de régression $\boldsymbol{\beta}$, et en utilisant les valeurs des variables auxiliaires pour les individus dans

et hors échantillon, les valeurs bootstrap de la variable expliquée sont générées pour tous les individus de la population, comme suit :

$$Y_{di}^{*} = \boldsymbol{x}_{di}'\widehat{\boldsymbol{\beta}} + u_d^{*} + e_{di}^{*}, \quad i = 1, \dots, N_d, d = 1, \dots, D.$$

Cela permet d'obtenir un recensement pour la variable expliquée, grâce auquel tout type d'indicateur pourra être calculé. Ce processus de génération est répété pour $a = 1, \dots, A$, ce qui permet d'obtenir A recensements complets. Pour chaque recensement a, on calcule l'indicateur d'intérêt $\delta_d^{*(a)} = \delta_d(\boldsymbol{y}_d^{*(a)})$, où $\boldsymbol{y}_d^{*(a)} = (Y_{d1}^{*(a)}, \dots, Y_{dN_d}^{*(a)})'$ sont les valeurs de la variable expliquée dans le domaine d pour le recensement bootstrap a. Au final, l'estimateur ELL s'obtient en faisant la moyenne sur les A recensements,

$$\hat{\delta}_d^{ELL} = \frac{1}{A}\sum_{a=1}^{A} \delta_d^{*(a)}.$$

De même, dans cette méthode, l'EQM est estimée comme suit :

$$\text{mse}_{ELL}(\hat{\delta}_d^{ELL}) = \frac{1}{A}\sum_{a=1}^{A} (\delta_d^{*(a)} - \hat{\delta}_d^{ELL})^2.$$

Pour estimer l'indicateur FGT d'ordre α utilisant cette méthode, nous écrivons d'abord cet indicateur comme une fonction des variables expliquées issues du modèle $Y_{di} = \log(E_{di} + c)$. En remplaçant $E_{di} = \exp(Y_{di}) - c$ dans la formule de l'indicateur FGT donnée en (1), nous obtenons:

$$F_{\alpha d} = \frac{1}{N_d}\sum_{i=1}^{N_d} \left(\frac{z + c - exp\,(Y_{di})}{z}\right)^{\alpha} I(\exp(Y_{di}) < z + c). \qquad (39)$$

Ainsi, nous calculons cet indicateur au moyen des valeurs Y_{di}^{*} générées pour chaque recensement a, comme suit :

$$F_{\alpha d}^{*(a)} = \frac{1}{N_d}\sum_{i=1}^{N_d} \left(\frac{z + c - exp\,(Y_{di}^{*(a)})}{z}\right)^{\alpha} I\left(\exp(Y_{di}^{*(a)}) < z + c\right),$$

et l'estimateur ELL de $F_{\alpha d}$ est ensuite approximé en faisant la moyenne de ces indicateurs sur les A recensements générés, i.e,

$$\hat{F}_{\alpha d}^{ELL} = \frac{1}{A}\sum_{a=1}^{A} F_{\alpha d}^{*(a)}.$$

Enfin, l'EQM de l'estimateur $\hat{F}_{\alpha d}^{ELL}$ est estimée comme suit :

$$\text{mse}_{ELL}(\hat{F}_{\alpha d}^{ELL}) = \frac{1}{A}\sum_{a=1}^{A} (F_{\alpha d}^{*(a)} - \hat{F}_{\alpha d}^{ELL})^2.$$

Il est facile de vérifier que, pour les domaines dont la taille de population N_d est grande (ce qui est le cas habituellement dans les applications réelles), si nous utilisons cette méthode pour estimer la moyenne sur un domaine d, $\bar{Y}_d$, en faisant la moyenne des $\bar{Y}_d^{*(a)} \approx \bar{\boldsymbol{X}}_d'\widehat{\boldsymbol{\beta}} + u_d^{*(a)}$ sur les A recensements, la moyenne des effets aléatoires bootstrap $u_d^{*(a)}$, prise sur les répétitions bootstrap, est $A^{-1}\sum_{a=1}^{A} u_d^{*(a)} \approx E(u_d) = 0$. Ainsi, l'estimateur ELL, $\hat{\bar{Y}}_d^{ELL} = E[\bar{Y}_d]$, tend vers l'estimateur synthétique par la régression,

$$\hat{\bar{Y}}_d^{ELL} = \bar{\boldsymbol{X}}_d{}'\hat{\boldsymbol{\beta}}.$$

Ceci est dû au fait que l'espérance marginale $E[\delta_d]$, sans prendre en compte le conditionnement des données disponibles de Y_{di} dans l'échantillon, n'utilise pas ces observations de l'échantillon et s'en tient donc à la prédiction obtenue par le modèle, sans considérer les effets aléatoires sur les domaines, car ceux-ci disparaissent. Ainsi, l'estimateur ELL présente les mêmes problèmes que l'estimateur synthétique par régression, à savoir qu'il peut être fortement biaisé si le modèle de régression sans les effets aléatoires n'est pas vérifié, c'est-à-dire si les variables auxiliaires considérées n'expliquent pas toute l'hétérogénéité de la variable expliquée entre les domaines.

De plus, dans la méthode de bootstrap utilisée, contrairement aux méthodes de bootstrap habituelles, le modèle n'est pas réajusté et estimé avec des échantillons bootstrap (qui devraient être tirés de recensements bootstrap). De ce fait, on ne reproduit pas le processus du monde réel dans le monde du bootstrap. Par conséquent, l'EQM estimée par cette méthode ne reproduit pas correctement l'erreur encourue dans l'estimation sur le monde réel. Enfin, dans la méthode ELL initiale, les effets aléatoires inclus dans le modèle concernent les grappes ou les unités d'échantillonnage du premier degré et non les domaines d'intérêt. Si l'on suit ce modèle, mais que les variables auxiliaires disponibles ne rendent pas compte de toute l'hétérogénéité entre les domaines, l'erreur de l'estimateur ELL peut être largement sous-estimée.

Résumé des caractéristiques de l'estimateur ELL :

Indicateurs cibles : paramètres généraux.

Données nécessaires :

- Microdonnées des p variables auxiliaires considérées, issues de la même enquête que celle où la variable d'intérêt est observée.
- Domaine d'intérêt obtenu à partir de la même enquête que celle où la variable d'intérêt est observée.
- Microdonnées des p variables auxiliaires considérées dans les domaines, issues d'un recensement ou de fichiers administratifs (mesurées de la même manière que dans l'enquête).

Avantages :

- Il est fondé sur des données individuelles, qui fournissent des informations plus détaillées que les données au niveau régional. De plus, la taille de l'échantillon est généralement beaucoup plus grande (n par rapport à D).
- Tous types d'indicateurs peuvent être estimés, tant qu'ils sont définis comme une fonction des variables expliquées Y_{di}.
- L'estimateur est sans biais sous le modèle si les paramètres du modèle sont connus.
- Une fois le modèle ajusté, l'estimateur peut être utilisé pour n'importe quelle sous-zone ou sous-domaine. Il peut même être estimé au niveau individuel.
- Une fois le modèle ajusté, tous les indicateurs souhaités (fonctions des Y_{di}) peuvent être estimés en même temps, sans qu'il soit nécessaire d'ajuster un modèle différent pour chaque indicateur.

Inconvénients :

- Les estimateurs ELL peuvent avoir une EQM élevée sous le modèle et peuvent même être moins performants que les estimateurs directs si l'hétérogénéité inexpliquée entre les zones

est importante, voir Molina et Rao (2010). Pour l'estimation des moyennes, les estimateurs ELL sont des estimateurs synthétiques par régression, qui supposent un modèle sans effets aléatoires sur les zones.

- Ils sont fondés sur un modèle. Il est donc nécessaire de vérifier que le modèle s'ajuste correctement sur les données.
- Ils ne sont pas sans biais sous le plan et peuvent avoir un biais considérable sous un plan informatif.
- Ils peuvent être sérieusement affectés par des valeurs aberrantes isolées.
- Si le modèle inclut des effets de grappe plutôt que des effets de zone d'intérêt, mais qu'il existe une hétérogénéité entre les zones, les estimateurs ELL sous-estiment la véritable EQM. Même si les effets de zone sont inclus dans le modèle, les estimateurs ELL de l'EQM n'estiment pas correctement la véritable EQM des estimateurs ELL pour chaque zone.

D. Meilleur prédicteur empirique sous le modèle avec erreurs emboîtées

Le meilleur prédicteur bayésien (BP[7]) fondé sur le modèle à erreurs emboîtées a été proposé par Molina et Rao (2010) pour estimer des indicateurs généraux non linéaires. Ces auteurs l'ont utilisé pour estimer le taux de la pauvreté et l'écart de pauvreté dans les provinces espagnoles par genre. Il a également été utilisé par le Conseil national pour l'évaluation de la politique de développement social (*Consejo Nacional para la Evaluación de la Política de Desarrollo Social (CONEVAL)*) au Mexique dans des études comparatives avec d'autres méthodes, comme l'ELL, pour l'estimation des indicateurs de pauvreté et d'inégalité dans les municipalités mexicaines. Cette méthode part du principe que les variables $Y_{di} = \log(E_{di} + c)$ suivent le modèle (32) sous l'hypothèse de normalité pour les effets aléatoires sur les domaines u_d et pour les erreurs e_{di}. Sous ce modèle, les vecteurs des variables pour chaque domaine, $\boldsymbol{y}_d = (Y_{d1}, \dots, Y_{dN_d})'$, $d = 1, \dots, D$, sont indépendants et vérifient $\boldsymbol{y}_d \sim^{ind} N(\boldsymbol{\mu}_d, \boldsymbol{V}_d)$, avec un vecteur espérance $\boldsymbol{\mu}_d = \boldsymbol{X}_d\boldsymbol{\beta}$, où $\boldsymbol{X}_d = (\boldsymbol{x}_{d1}, \dots, \boldsymbol{x}_{dN_d})'$, et une matrice de variance-covariance $\boldsymbol{V}_d = \sigma_u^2 \mathbf{1}_{N_d} \mathbf{1}_{N_d}{}' + \sigma_e^2 \boldsymbol{A}_d$, où $\boldsymbol{A}_d = \mathrm{diag}(k_{di}^2; i = 1, \dots, N_d)$. Pour un indicateur général défini comme une fonction des $\boldsymbol{y}_d$, i.e., $\delta_d = \delta_d(\boldsymbol{y}_d)$, le meilleur prédicteur est celui qui minimise l'EQM et il a pour expression:

$$\tilde{\delta}_d^B(\boldsymbol{\theta}) = E_{\boldsymbol{y}_{dr}}[\delta_d(\boldsymbol{y}_d)|\boldsymbol{y}_{ds}; \boldsymbol{\theta}], \tag{40}$$

où l'espérance est prise par rapport à la distribution du vecteur des valeurs hors échantillon $\boldsymbol{y}_{dr}$ pour le domaine d, conditionnellement aux valeurs dans l'échantillon $\boldsymbol{y}_{ds}$. Cette distribution conditionnelle dépend de la vraie valeur des paramètres du modèle, soit $\boldsymbol{\theta}$. En remplaçant $\boldsymbol{\theta}$ par un estimateur convergent $\widehat{\boldsymbol{\theta}}$ dans la formule du meilleur prédicteur (40), on obtient ce que l'on nomme le meilleur estimateur empirique bayésien (EB), $\hat{\delta}_d^{EB} = \tilde{\delta}_d^B(\widehat{\boldsymbol{\theta}})$. À nouveau, les méthodes d'estimation usuelles, qui fournissent des estimateurs convergents même en l'absence de normalité, sont les méthodes ML et REML, les deux sous une vraisemblance normale, et la méthode Henderson III.

Sous le modèle à erreurs emboîtées (32), la distribution de $\boldsymbol{y}_{dr}|\boldsymbol{y}_{ds}$, requise pour calculer le meilleur prédicteur (40), s'obtient comme suit. Tout d'abord, nous décomposons les matrices $\boldsymbol{X}_d$ et $\boldsymbol{V}_d$ en blocs relatifs aux données dans et hors échantillon, de la même manière que l'on a décomposé $\boldsymbol{y}_d$, i.e,

[7] Best/Bayes predictor.

$$\boldsymbol{y}_d = \begin{pmatrix} \boldsymbol{y}_{ds} \\ \boldsymbol{y}_{dr} \end{pmatrix}, \quad \boldsymbol{X}_d = \begin{pmatrix} \boldsymbol{X}_{ds} \\ \boldsymbol{X}_{dr} \end{pmatrix}, \quad \boldsymbol{V}_d = \begin{pmatrix} \boldsymbol{V}_{ds} & \boldsymbol{V}_{dsr} \\ \boldsymbol{V}_{drs} & \boldsymbol{V}_{dr} \end{pmatrix}.$$

Puisque $\boldsymbol{y}_d$ suit une distribution normale, les distributions conditionnelles sont également normales, i.e,

$$\boldsymbol{y}_{dr}|\boldsymbol{y}_{ds} \overset{ind}{\sim} N\left(\boldsymbol{\mu}_{dr|s}, \boldsymbol{V}_{dr|s}\right), \quad d = 1, \dots, D, \tag{41}$$

où le vecteur des espérances conditionnelles et la matrice de variance-covariance correspondante prennent la forme :

$$\boldsymbol{\mu}_{dr|s} = \boldsymbol{X}_{dr}\boldsymbol{\beta} + \gamma_d(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}^T\boldsymbol{\beta})\mathbf{1}_{N_d - n_d}, \tag{42}$$

$$\boldsymbol{V}_{dr|s} = \sigma_u^2(1-\gamma_d)\mathbf{1}_{N_d-n_d}\mathbf{1}_{N_d-n_d}^T + \sigma_e^2 \mathrm{diag}_{i\in r_d}(k_{di}^2), \tag{43}$$

où $\mathbf{1}_k$ est le vecteur de 1 de taille k. Spécifiquement, pour l'individu $i \in r_d$, on a :

$$Y_{di}|\boldsymbol{y}_{ds} \sim N(\mu_{di|s}, \sigma_{di|s}^2), \tag{44}$$

où l'espérance conditionnelle et la variance s'expriment sous la forme

$$\mu_{di|s} = \boldsymbol{x}_{di}{}'\boldsymbol{\beta} + \gamma_d(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}^T\boldsymbol{\beta}), \tag{45}$$

$$\sigma_{di|s}^2 = \sigma_u^2(1-\gamma_d) + \sigma_e^2 k_{di}^2. \tag{46}$$

Si nous souhaitons maintenant estimer l'indicateur de pauvreté FGT d'ordre α, $\delta_d = F_{\alpha d}$, nous faisons tout d'abord l'hypothèse que $Y_{di} = \log(E_{di} + c)$, pour $c > 0$, vérifie le modèle avec erreurs emboîtées. Nous récrivons l'indicateur FGT en question comme fonction des variables expliquées dans le modèle Y_{di}, i.e., comme en (39), et nous calculons l'espérance qui définit le meilleur prédicteur $\tilde{F}_{\alpha d}^B = E_{\boldsymbol{y}_{dr}}[F_{\alpha d}|\boldsymbol{y}_{ds}; \boldsymbol{\theta}]$. Pour cela, nous séparons la somme qui définit l'indicateur FGT donné en (1) en deux parties, relatives aux données dans et hors échantillon, et, en introduisant l'espérance dans la somme, nous obtenons :

$$\tilde{F}_{\alpha d}^B(\boldsymbol{\theta}) = \frac{1}{N_d}\left(\sum_{i\in s_d} F_{\alpha,di} + \sum_{i\in r_d} \tilde{F}_{\alpha,di}^B(\boldsymbol{\theta})\right), \tag{47}$$

où $\tilde{F}_{\alpha,di}^B(\boldsymbol{\theta}) = E[F_{\alpha,di}|\boldsymbol{y}_{ds}; \boldsymbol{\theta}]$ et l'espérance est prise par rapport à la distribution de $Y_{di}|\boldsymbol{y}_{ds}$, $i \in r_d$, donnée en (44)-(46). Pour $\alpha = 0{,}1$, les espérances sont faciles à calculer, et s'expriment respectivement comme

$$\tilde{F}_{0,di}^B(\boldsymbol{\theta}) = \Phi(\alpha_{di}), \tag{48}$$

$$\tilde{F}_{1,di}^B(\boldsymbol{\theta}) = \Phi(\alpha_{di})\left\{1 - \frac{1}{z}\left[\exp\left(\mu_{di|s} + \frac{\sigma_{di|s}^2}{2}\right)\frac{\Phi(\alpha_{di} - \sigma_{di|s})}{\Phi(\alpha_{di})} - c\right]\right\}, \tag{49}$$

où $\Phi(\cdot)$ est la fonction de répartition d'une variable aléatoire normale standard et

$$\alpha_{di} = [\log(z + c) - \mu_{di|s}]/\sigma_{di|s}, \text{ avec } \mu_{di|s} \text{ et } \sigma_{di|s}^2 \text{ donnés en (45)-(46).}$$

Pour des indicateurs plus complexes $\delta_d = \delta_d(\boldsymbol{y}_d)$, par exemple les indicateurs FGTpour $\alpha \neq 0{,}1$, l'espérance définissant le meilleur prédicteur peut souvent ne pas être calculable analytiquement. Dans ces cas, le meilleur prédicteur peut être approché empiriquement en utilisant une simulation de Monte Carlo. Le processus serait alors le suivant :

- Obtenir un estimateur $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}', \hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ pour le vecteur des paramètres $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2)'$ en ajustant le modèle (32) aux données $(\boldsymbol{y}_s, \boldsymbol{X}_s)$.
- Générer, pour $a = 1, \dots, A$, des vecteurs des variables expliquées pour les unités hors

échantillon du domaine d, $\boldsymbol{y}_{dr}^{(a)}$, fondés sur la distribution de $\boldsymbol{y}_{dr}|\boldsymbol{y}_{ds}$ donnée en (41)-(43), où $\boldsymbol{\theta}$ est remplacé par son estimateur $\widehat{\boldsymbol{\theta}}$ obtenu en (a).

- Fusionner le vecteur généré $\boldsymbol{y}_{dr}^{(a)}$ avec celui des données de l'échantillon $\boldsymbol{y}_{ds}$ de façon à former un vecteur de recensement pour le domaine d, $\boldsymbol{y}_d^{(a)} = (\boldsymbol{y}_{ds}', (\boldsymbol{y}_{dr}^{(a)})')'$. En utilisant $\boldsymbol{y}_d^{(a)}$, calculer l'indicateur d'intérêt $\delta_d^{(a)} = \delta_d(\boldsymbol{y}_d^{(a)})$ et répéter pour $a = 1, \dots, A$. L'approximation de Monte Carlo du prédicteur EB de l'indicateur δ_d est obtenue en prenant la moyenne des indicateurs sur les A recensements simulés, i.e.

$$\hat{\delta}_d^{EB} = \frac{1}{A}\sum_{a=1}^{A} \delta_d^{(a)}. \tag{50}$$

À l'étape (b), on doit simuler A fois un vecteur $\boldsymbol{y}_{dr}^{(a)}$ ayant une distribution normale multivariée de taille $N_d - n_d$, qui peut être réellement conséquente (par exemple, de la taille d'une province), ce qui peut être très difficile à mettre en oeuvre, voire impossible, en raison de la grande taille du vecteur multivarié à générer. On peut éviter ceci en notant que la matrice de variance-covariance de ce vecteur, $\boldsymbol{V}_{dr|s}$, donnée en (43), correspond à celle d'un vecteur aléatoire $\boldsymbol{y}_{dr}^{(a)}$ généré à partir du modèle

$$\boldsymbol{y}_{dr}^{(a)} = \boldsymbol{\mu}_{dr|s} + v_d^{(a)}\mathbf{1}_{N_d-n_d} + \boldsymbol{\epsilon}_{dr}^{(a)}, \tag{51}$$

où $v_d^{(a)}$ et $\boldsymbol{\epsilon}_{dr}^{(a)}$ sont indépendants et vérifient, respectivement

$$v_d^{(a)} \sim N(0, \sigma_u^2(1-\gamma_d)), \quad \boldsymbol{\epsilon}_{dr}^{(a)} \sim N(\mathbf{0}_{N_d-n_d}, \sigma_e^2 \mathrm{diag}_{i\in r_d}(k_{di}^2)); \tag{52}$$

(voir Molina et Rao (2010)). En utilisant le modèle (51)-(52), au lieu de générer un vecteur normal multivarié $\boldsymbol{y}_{dr}^{(a)}$ de taille $N_d - n_d$, il est seulement nécessaire de générer les $1 + N_d - n_d$ variables normales indépendantes $v_d^{(a)} \sim^{ind} N(0, \sigma_u^2(1-\gamma_d))$ et $\epsilon_{di}^{(a)} \sim^{ind} N(0, \sigma_e^2 k_{di}^2)$, pour $i \in r_d$. En utilisant le vecteur $\boldsymbol{y}_{dr}^{(a)}$ généré à partir du modèle (51), à l'étape (c) nous construisons le vecteur de recensement $\boldsymbol{y}_d^{(a)} = (\boldsymbol{y}_{ds}', (\boldsymbol{y}_{dr}^{(a)})')'$ et nous calculons l'indicateur d'intérêt $\delta_d^{(a)} = \delta_d(\boldsymbol{y}_d^{(a)})$.

Pour un domaine non échantillonné d (i.e. avec $n_d = 0$), nous générons $\boldsymbol{y}_{dr}^{(a)}$ à partir du modèle (51) en prenant $\gamma_d = 0$ et, comme il n'y a pas de partie échantillonnée dans ce cas, le vecteur de recensement du domaine d est égal au vecteur généré $\boldsymbol{y}_d^{(a)} = \boldsymbol{y}_{dr}^{(a)}$.

Dans le cas d'indicateurs complexes, il est compliqué de calculer des approximations analytiques pour l'EQM des prédicteurs EB correspondants. Molina et Rao (2010) décrivent une méthode de bootstrap paramétrique pour estimer l'EQM fondée sur la méthode de bootstrap pour les populations finies présentée par González-Manteiga et al. (2008). Cette méthode consiste en les étapes suivantes :

- Ajuster le modèle (32) aux données de l'échantillon $\boldsymbol{y}_s = (\boldsymbol{y}_{1s}', \dots, \boldsymbol{y}_{Ds}')'$, pour obtenir des estimations des paramètres du modèle, $\widehat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ et $\hat{\sigma}_e^2$.
- Générer des effets bootstrap sur les domaines comme suit :
$$u_d^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2), \quad d = 1, \dots, D.$$
- Générer, indépendamment de $u_1^{*(b)}, \dots, u_D^{*(b)}$, des erreurs bootstrap
$$e_{di}^{*(b)} \sim^{iid} N(0, \hat{\sigma}_e^2), \quad i = 1, \dots, N_d, d = 1, \dots, D.$$
- Générer une population bootstrap (ou recensement) des valeurs de la variable expliquée au moyen du modèle,
$$Y_{di}^{*(b)} = \boldsymbol{x}_{di}'\widehat{\boldsymbol{\beta}} + u_d^{*(b)} + e_{di}^{*(b)}, \quad i = 1, \dots, N_d, d = 1, \dots, D.$$

- On définit le vecteur de recensement des variables d'intérêt du domaine d, donné par $\boldsymbol{y}_d^{*(b)} = (Y_{d1}^{*(b)}, \dots, Y_{dN_d}^{*(b)})'$. Calculer les indicateurs d'intérêt à partir du recensement bootstrap.
$$\delta_d^{*(b)} = \delta_d(\boldsymbol{y}_d^{*(b)}), d = 1, \dots, D.$$
- Pour l'échantillon d'origine $s = s_1 \cup \cdots \cup s_D$, soit $\boldsymbol{y}_s^{*(b)} = ((\boldsymbol{y}_{1s}^{*(b)})', \dots, (\boldsymbol{y}_{Ds}^{*(b)})')'$ le vecteur contenant les observations bootstrap dont les indices sont dans l'échantillon, i.e. contenant les variables $Y_{di}^{*(b)}$, $i \in s_d$, $d = 1, \dots, D$. À nouveau, ajuster le modèle (32) aux données bootstrap $\boldsymbol{y}_s^{*(b)}$ et obtenir les prédicteurs bootstrap EB des indicateurs d'intérêt, $\hat{\delta}_d^{EB*(b)}$, $d = 1, \dots, D$.
- Répéter les étapes 2) - 6) pour $b = 1, \dots, B$, et l'on obtient alors les vraies valeurs, $\delta_d^{*(b)}$, et les prédicteurs EB correspondants, $\hat{\delta}_d^{EB*(b)}$, pour chaque domaine $d = 1, \dots, D$, et pour chaque réplication bootstrap, $b = 1, \dots, B$.
- Les estimateurs naïfs bootstrap des EQM des prédicteurs EB, $\hat{\delta}_d^{EB}$, ont pour expression
$$\text{mse}_B(\hat{\delta}_d^{EB}) = B^{-1} \sum_{b=1}^{B} \left(\hat{\delta}_d^{EB*(b)} - \delta_d^{*(b)}\right)^2, \quad d = 1, \dots, D.$$

Notons que, pour estimer des indicateurs complexes, la méthode ELL décrite dans le chapitre précédent et la méthode EB présentée dans ce chapitre nécessitent des données provenant d'une enquête avec des observations de la variable d'intérêt et des variables auxiliaires pour tous les domaines, $\{(y_{di}, \boldsymbol{x}_{di}); i \in s_d, d = 1, \dots, D\}$, ainsi qu'un recensement avec les valeurs des mêmes variables auxiliaires pour toutes les unités de la population, $\{\boldsymbol{x}_{di}; i = 1, \dots, N_d, d = 1, \dots, D\}$.

En principe, la méthode EB doit également identifier dans le recensement les unités qui sont aussi dans l'échantillon relatif à chaque domaine, s_d. Dans la pratique, il n'est pas toujours possible de relier les données de l'enquête et du recensement. Cependant, la taille de l'échantillon du domaine, n_d, est généralement très petite par rapport à la taille de la population du domaine, N_d. Ensuite, nous pouvons utiliser le meilleur prédicteur du recensement proposé par Correa, Molina et Rao (2012), qui est obtenu en calculant les espérances conditionnelles $\tilde{F}_{\alpha,di}^B(\boldsymbol{\theta})$, y compris pour les individus de l'échantillon comme s'ils n'étaient pas observés, c'est-à-dire que le meilleur prédicteur du recensement de $F_{\alpha d}$ a pour expression:

$$\tilde{F}_{\alpha d}^{CB}(\boldsymbol{\theta}) = \frac{1}{N_d} \sum_{i=1}^{N_d} \tilde{F}_{\alpha,di}^B(\boldsymbol{\theta}). \tag{53}$$

De la même façon que le prédicteur EB, nous définissons le prédicteur EB de recensement de $F_{\alpha d}$, en remplaçant dans (53) $\boldsymbol{\theta}$ par un estimateur convergent. Si l'espérance définissant $\tilde{F}_{\alpha,di}^B(\boldsymbol{\theta})$ ne peut pas être calculée analytiquement, comme cela se produit lorsque l'indicateur a une forme compliquée, dans chaque réplication de la procédure de Monte Carlo décrite dans (1)-(3), nous générons le vecteur de recensement complet $\boldsymbol{y}_d$ au lieu du seul vecteur d'observations hors échantillon $\boldsymbol{y}_{dr}$; c'est-à-dire que nous appliquons l'approximation de Monte Carlo (50) en générant $\boldsymbol{y}_d^{(a)} = \boldsymbol{\mu}_{d|s} + v_d^{(a)} \mathbf{1}_{N_d - n_d} + \boldsymbol{\epsilon}_d^{(a)}$,où $\boldsymbol{\mu}_{d|s} = \boldsymbol{X}_d \boldsymbol{\beta} + \gamma_d(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}^T \boldsymbol{\beta}) \mathbf{1}_{N_d}$,et $\boldsymbol{\epsilon}_d^{(a)} \sim N(\mathbf{0}_{N_d}, \sigma_e^2 \text{diag}_{i=1,\dots,N_d}(k_{di}^2))$. Si la fraction d'échantillonnage n_d/N_d est négligeable, comme c'est le cas dans la plupart des cas réels, l'estimateur EB de recensement (EB Census) de $\delta_d = F_{\alpha d}$ sera pratiquement égal à l'estimateur EB d'origine.

Pour les indicateurs dont le calcul a un coût élevé, comme ceux qui nécessitent de classer les individus de la population en fonction de leur pouvoir d'achat, comme les indicateurs "*Fuzzy monetary*" et les "*Fuzzy supplementary indicators*", le temps de calcul de la procédure totale, y compris la méthode bootstrap pour le calcul de l'EQM, augmente. Dans ce cas, Ferretti et Molina (2012) ont proposé une variante du prédicteur EB, connue sous le nom de "fast EB", qui est beaucoup plus rapide en termes de calcul. Dans la procédure de Monte Carlo (1)-(3) pour l'approximation du prédicteur EB, cette procédure

remplace la génération du recensement à l'étape (2) par la génération d'un échantillon (différent dans chaque réplication de Monte Carlo) et le calcul des valeurs réelles des indicateurs à l'étape (3) par le calcul d'estimateurs basés sur le plan de sondage, qui ne nécessitent qu'un échantillon au lieu du recensement complet.

Propriétés du prédicteur EB (approximé par l'estimateur EB de recensement si n_d/N_d est négligeable) :

Indicateurs cibles : paramètres généraux.

Données nécessaires :

- Microdonnées des p variables auxiliaires considérées, provenant de la même enquête que celle où la variable d'intérêt est observée.
- Domaine d'intérêt obtenu à partir de la même enquête que celle où la variable d'intérêt est observée.
- Microdonnées des p variables auxiliaires considérées provenant d'un recensement ou de fichiers administratifs (mesurées de la même manière que dans l'enquête).

Avantages:

- Ces prédicteurs sont fondés sur des données au niveau individuel, qui fournissent des informations plus détaillées que les données au niveau du domaine (il est également possible d'incorporer des variables au niveau du domaine). En outre, la taille de l'échantillon est généralement beaucoup plus importante (n par rapport à D).
- Tous les indicateurs peuvent être estimés, à partir du moment où ils sont définis comme une fonction des variables expliquées Y_{di}.
- Ils sont sans biais sous le modèle si les paramètres du modèle sont connus.
- Ils sont optimaux dans le sens où ils minimisent l'EQM sous le modèle, pour des valeurs connues des paramètres.
- Ils sont nettement plus performants que les estimateurs ELL en termes d'EQM sous le modèle (32) lorsque l'hétérogénéité inexpliquée entre les zones est significative. Pour les domaines non échantillonnés (avec $n_d = 0$), les estimateurs EB et ELL sont pratiquement identiques. Ils seront également pratiquement les mêmes, dans ce cas pour tous les domaines, si toute l'hétérogénéité entre les domaines est expliquée par les variables auxiliaires ($\sigma_u^2 = 0$).
- Une fois le modèle ajusté, il peut être estimé pour n'importe quelle sous-zone ou sous-domaine. Il peut même être estimé au niveau individuel.
- Une fois le modèle ajusté, tous les indicateurs souhaités (qui sont une fonction de Y_{di}) peuvent être estimés en même temps, sans qu'il soit nécessaire d'ajuster un modèle différent pour chaque indicateur.

Inconvénients :

- Ils sont fondés sur un modèle. Il est donc nécessaire de vérifier que le modèle s'ajuste correctement (par exemple, à l'aide des résidus).
- Ils ne prennent pas en compte le plan d'échantillonnage. Ils ne sont pas sans biais sous le plan de sondage et peuvent présenter un biais considérable sous un plan informatif.
- Ils peuvent être sérieusement affectés par des valeurs aberrantes isolées ou l'absence de normalité.

- Les estimateurs de l'EQM obtenus à l'aide de la méthode bootstrap paramétrique sont exigeants en termes de calcul.

Exemple 7. Estimateurs EB du taux de pauvreté, avec R. Poursuivant les exemples précédents, nous montrons comment obtenir les estimateurs EB du taux de pauvreté avec R, sur la base d'un modèle à erreurs imboîtées pour le logarithme du revenu (transformé avec une constante). Le seuil de pauvreté a été établi à l'avance comme étant égal à 60% du revenu médian, et il s'avère être $z = 6557.143$. En utilisant ce seuil, nous devons définir la fonction qui nous donne le taux de pauvreté :

```
povertyincidence <- function(y) {
 result <- mean(y < 6557.143)
 return (result)
}
```

Nous faisons maintenant appel à la fonction qui calcule les estimateurs EB en sélectionnant la fonction de taux de pauvreté comme indicateur, en prenant la transformation logarithmique, et en ajoutant la constante=3500 au revenu avant cette transformation, et en utilisant des réplications pour l'approximation de Monte Carlo des estimateurs EB. La constante mentionnée ci-dessus est choisie de manière à ce que les résidus de l'ajustement présentent une distribution approximativement symétrique, puisque la méthode EB décrite est fondée sur une distribution normale. Avant d'appeler la fonction, nous définissons les racines du générateur de nombres aléatoires afin que la fonction nous donne les mêmes estimations si l'on répète l'appel à cette fonction, et nous initialisons le vecteur qui contiendra les estimateurs EB :

```
povinc.EB<-numeric(D)
set.seed(123)      # Nous fixons les racines du générateur aléatoire
res.EB<-ebBHF(income~age2+age3+age4+age5+nat1+educ1+educ3+labor1+labor2,dom=prov,

selectdom=selprov,Xnonsample=Xoutsamp,MC=50,constant=3500,indicator=povertyincidence)
povinc.EB[selprov]<-res.EB$eb$eb$eb
```

Quel que soit le modèle, les résidus doivent être analysés afin de vérifier que les données ne présentent pas d'indications contraires à la spécification du modèle supposé. Comme la méthode EB exige la normalité, nous traçons un histogramme et un graphique q-q de normalité des résidus :

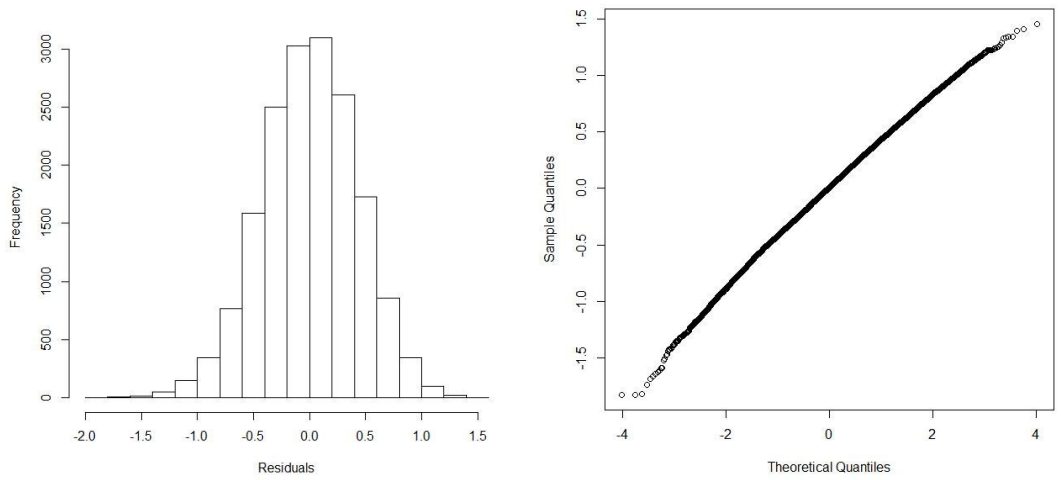
```
resid.EB<-res.EB$fit$residuals
hist(resid.EB,main="",xlab="Residuals")
qqnorm(resid.EB,main="")
```

Les deux graphiques (figure 8) montrent que la distribution des résidus est approximativement normale. En revanche, si nous ajustons le modèle au revenu sans la transformation logarithmique, l'histogramme et le graphique de normalité q-q (non reproduits ici pour ne pas alourdir le texte) montrent une distribution nettement asymétrique vers la droite. Cette transformation est donc nécessaire pour ne pas s'éloigner de l'hypothèse de normalité.

Enfin, nous calculons les estimateurs bootstrap de l'EQM des estimateurs EB avec B=200 réplications bootstrap et MC=50 réplications pour l'approximation de Monte Carlo des estimateurs EB.

```
set.seed(123)
povinc.mse.res<-
pbmseebBHF(income~age2+age3+age4+age5+nat1+educ1+educ3+labor1+labor2,dom=prov,selectdom
=selprov,Xnonsample=Xoutsamp,B=200,MC=50,constant=3500,
 indicator=povertyincidence)
povinc.eb.mse<-numeric(D)
povinc.eb.mse[selprov]<-povinc.mse.res$mse$mse
```

Figure 8
Histogramme (à gauche) et graphique q-q de normalité (à droite) des résidus de l'ajustement du modèle avec erreurs emboîtées pour le logarithme du revenu
(En unités)



Source : D'après l'auteur.

Finalement, nous comparons graphiquement les estimateurs EB avec les estimateurs directs du taux de pauvreté HT, FH et EBLUP, fondés sur le modèle avec erreurs emboîtées, pour les provinces sélectionnées :

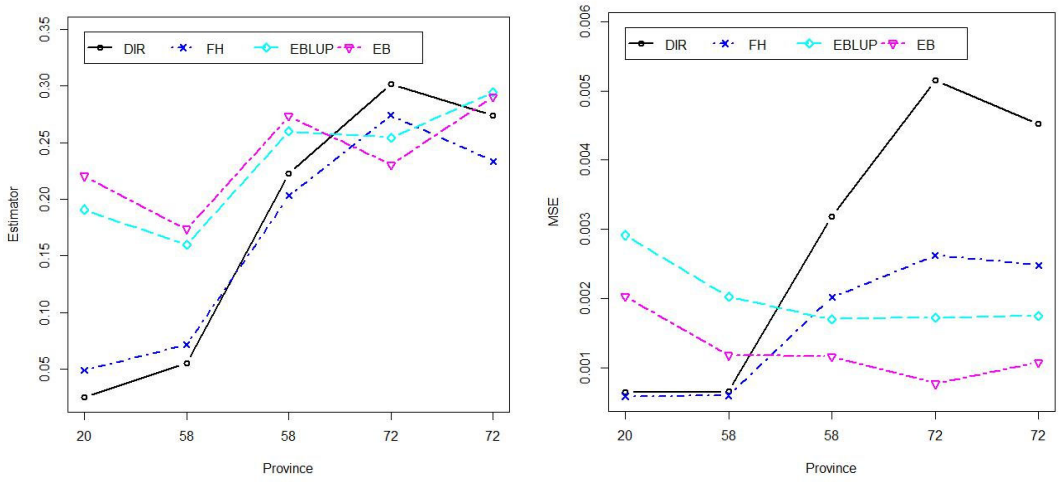
```
k<-6

M<-max(povinc.dir[selprov],povinc.FH [selprov],povinc.BHF [selprov],povinc.EB [selprov])
m<-min(povinc.dir[selprov],povinc.FH [selprov],povinc.BHF [selprov],povinc.EB [selprov])
plot(1:5,povinc.dir[selprov],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="Estimator",
 xaxt="n")
points(1:5,povinc.dir[selprov],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:5,povinc.FH[selprov],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:5,povinc.BHF[selprov],type="b",col=5,lty=5,pch=5,lwd=2)
points(1:5,povinc.EB[selprov],type="b",col=6,lty=6,pch=6,lwd=2)
axis(1, at=1:5, labels=nd[selprov])
legend(1,M+(M-m)/k,legend=c("DIR", "FH", "EBLUP", "EB"),ncol=4,col=c(1,4,5,6),lwd=rep(2,4),
 lty=c(1,4,5.6),pch=c(1,4,5.6))

M<-max(povinc.dir.var[selprov],povinc.FH.mse[selprov],povinc.BHF.mse[selprov],
 povinc.eb.mse[selprov])
m<-min(povinc.dir.var[selprov],povinc.FH.mse[selprov],povinc.BHF.mse[selprov],
 povinc.eb.mse[selprov])
plot(1:5,povinc.dir.var[selprov],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="CV",
 xaxt="n")
points(1:5,povinc.dir.var[selprov],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:5,povinc.FH.mse[selprov],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:5,povinc.BHF.mse[selprov],type="b",col=5,lty=5,pch=5,lwd=2)
points(1:5,povinc.eb.mse[selprov],type="b",col=6,lty=6,pch=6,lwd=2)
axis(1, at=1:5, labels=nd[selprov])
legend(1,M+(M-m)/k,legend=c("DIR", "FH", "EBLUP", "EB"),ncol=4,col=c(1,4,5,6),lwd=rep(2,4),
```

lty=c(1,4,5.6),pch=c(1,4,5.6))

Selon la figure 9 (à gauche), les estimateurs EB sont très similaires aux EBLUP. Cela paraît raisonnable puisque les deux sont fondés sur un modèle au niveau individuel, bien que les estimateurs EB s'ajustent sur le modèle avec le logarithme du revenu, alors que les EBLUP s'ajustent sur celui avec un indicateur binaire traduisant le fait d'avoir ou non un revenu inférieur au seuil (variable "pauvre").

Théoriquement, le modèle supposé par les EBLUP n'est pas exact, puisque la variable expliquée est binaire et que les prédicteurs peuvent fournir des valeurs en dehors de l'intervalle. De plus, malgré la similarité entre les estimations EB et EBLUP, la figure 9 (à droite) indique que les estimateurs EB sont plus efficaces que les EBLUP.

Figure 9
Estimations EB et EBLUP fondées sur un modèle à erreurs emboîtées, FH, et HT direct (à gauche), et EQM de ces estimateurs (à droite) pour les provinces sélectionnées
(En proportions)



Source : D'après l'auteur.

E. Méthode bayésienne hiérarchique sous le modèle à erreurs emboîtées

Le calcul des estimateurs EB (ou EB-Recensement) ainsi que de leurs EQM estimées est très exigeant sur le plan informatique et peut ne pas être opérationnel pour de très grandes populations ou pour des indicateurs très complexes (par exemple, ceux qui nécessitent une structure d'ordre). Notons que pour obtenir l'approximation de Monte Carlo de l'estimateur EB, il est nécessaire de construire A recensements $\boldsymbol{y}^{(a)}$, $a = 1, \dots, A$, qui peuvent être très volumineux. De plus, lorsqu'on utilise le bootstrap pour estimer l'EQM, l'approximation de Monte Carlo doit être répétée pour chaque réplication du bootstrap. Afin de développer une méthode plus efficace en termes de calcul, Molina, Nandram et Rao (2014) ont proposé la méthode bayésienne hiérarchique (HB[8]) pour estimer des indicateurs généraux. Cette procédure ne nécessite pas l'utilisation de méthodes de bootstrap pour estimer l'EQM, car elle fournit des échantillons de la distribution a posteriori, à partir desquels les variances a posteriori qui jouent le rôle de l'EQM, ou toute autre mesure synthétique, peuvent être facilement obtenues.

La méthode HB est fondée sur une reparamétrisation du modèle avec erreurs emboîtées (32) en termes de coefficient de corrélation intraclasse $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$ et la prise en compte des

[8] Hierarchical Bayes.

distributions a priori pour les paramètres du modèle $(\boldsymbol{\beta}, \rho, \sigma_e^2)$ qui reflètent le manque d'information a priori sur ceux-ci. Plus précisément; nous considérons le modèle HB suivant :

$$\text{(i) } Y_{di}|u_d, \boldsymbol{\beta}, \sigma_e^2 \overset{ind}{\sim} N\,(\boldsymbol{x}_{di}'\boldsymbol{\beta} + u_d, \sigma_e^2 k_{di}^2), \quad i = 1, \dots, N_d,$$
$$\text{(ii) } u_d|\rho, \sigma_e^2 \overset{iid}{\sim} N\left(0, \frac{\rho}{1-\rho}\sigma_e^2\right), \quad d = 1, \dots, D,$$
$$\text{(iii) } \pi(\boldsymbol{\beta}, \rho, \sigma_e^2) \propto \frac{1}{\sigma_e^2}, \quad \epsilon \leq \rho \leq 1 - \epsilon,\ \sigma_e^2 > 0, \boldsymbol{\beta} \in R^p,$$

où $\epsilon > 0$ est choisi très petit pour refléter le manque d'information a priori. (Voir l'application proposée par Molina, Nandram, et Rao (2014), où l'inférence n'est pas sensible à de petites variations de ϵ.).

La distribution a posteriori des paramètres du modèle peut être calculée sur la base des distributions conditionnelles en utilisant les enchaînements suivants. Tout d'abord, notons que, dans la méthode HB, les effets aléatoires $\boldsymbol{u} = (u_1, \dots, u_D)'$ sont considérés comme des paramètres additionnels. Ensuite, la densité jointe du vecteur des paramètres $\boldsymbol{\theta} = (\boldsymbol{u}', \boldsymbol{\beta}', \sigma_e^2, \rho)'$ conditionnellement aux observations de l'échantillon $\boldsymbol{y}_s$ s'écrit

$$\pi(\boldsymbol{u}, \boldsymbol{\beta}, \sigma_e^2, \rho|\boldsymbol{y}_s) = \pi_1(\boldsymbol{u}|\boldsymbol{\beta}, \sigma_e^2, \rho, \boldsymbol{y}_s)\pi_2(\boldsymbol{\beta}|\sigma_e^2, \rho, \boldsymbol{y}_s)\pi_3(\sigma_e^2|\rho, \boldsymbol{y}_s)\pi_4(\rho|\boldsymbol{y}_s), \tag{54}$$

où toutes les densités conditionnelles à l'exception de π_4 ont des formes connues. Comme ρ est défini dans un intervalle fermé à l'intérieur de (0,1), nous pouvons générer des valeurs de π_4 en utilisant une méthode par pas. Pour plus de détails voir Molina, Nandram et Rao (2014)). Ainsi, des échantillons de $\boldsymbol{\theta} = (\boldsymbol{u}', \boldsymbol{\beta}', \sigma_e^2, \rho)'$ peuvent être générés directement à partir de la distribution a posteriori donnée en (54), sans qu'il soit besoin d'utiliser de méthodes de type Chaîne de Markov – Monte Carlo (MCMC). Sous des conditions assez générales, l'obtention d'une distribution a posteriori indépendante peut être assurée.

Etant donné $\boldsymbol{\theta}$, sous le modèle HB (i)-(iii), les variables Y_{di} pour tous les individus de la population sont indépendantes et vérifient

$$Y_{di}|\boldsymbol{\theta} \overset{ind}{\sim} N\,(\boldsymbol{x}_{di}'\boldsymbol{\beta} + u_d, \sigma_e^2 k_{di}^2), \quad i = 1, \dots, N_d,\ d = 1, \dots, D. \tag{55}$$

La densité conditionnelle de $\boldsymbol{y}_{dr}$ a pour expression

$$f(\boldsymbol{y}_{dr}|\boldsymbol{y}_s) = \int \prod_{i \in r_d} f\,(Y_{di}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y}_s)d\boldsymbol{\theta},$$

où $\pi(\boldsymbol{\theta}|\boldsymbol{y}_s)$ est donné en (54). Finalement, l'estimateur HB du paramètre $\delta_d = \delta_d(\boldsymbol{y}_d)$ est

$$\hat{\delta}_d^{HB} = E_{\boldsymbol{y}_{dr}}(\delta_d|\boldsymbol{y}_s) = \int \delta_d\,(\boldsymbol{y}_d) f(\boldsymbol{y}_{dr}|\boldsymbol{y}_s)d\boldsymbol{y}_{dr}. \tag{56}$$

Cet estimateur peut être approximé en utilisant une simulation de Monte Carlo. Pour ce faire, nous générons des échantillons de la distribution a posteriori $\pi(\boldsymbol{\theta}|\boldsymbol{y}_s)$ comme suit. Tout d'abord, nous générons une valeur $\rho^{(a)}$ à partir de $\pi_4(\rho|\boldsymbol{y}_s)$ en utilisant la méthode par pas (voir Molina, Nandram, & Rao, 2014) ; puis, nous générons $\sigma_e^{2(a)}$ à partir de $\pi_3\left(\sigma_e^2\middle|\rho^{(a)}, \boldsymbol{y}_s\right)$; ensuite, $\boldsymbol{\beta}^{(a)}$ est généré à partir de $\pi_2(\boldsymbol{\beta}|\sigma_e^{2(a)}, \rho^{(a)}, \boldsymbol{y}_s)$ et, finalement, $\boldsymbol{u}^{(a)}$ est généré à partir de $\pi_1(\boldsymbol{u}|\boldsymbol{\beta}^{(a)}, \sigma_e^{2(a)}, \rho^{(a)}, \boldsymbol{y}_s)$. Ce processus est répété A fois, de façon à obtenir un échantillon aléatoire $\boldsymbol{\theta}^{(a)}$, $a = 1, \dots, A$, à partir de $\pi(\boldsymbol{\theta}|\boldsymbol{y}_s)$. Pour chaque valeur générée $\boldsymbol{\theta}^{(a)}$ de $\pi(\boldsymbol{\theta}|\boldsymbol{y}_s)$, nous générons les valeurs hors échantillon $\{Y_{di}^{(a)}, i \in r_d\}$ à partir de la distribution donnée en (55) ce qui permet d'obtenir, pour chaque domaine d, le vecteur des variables hors échantillon $\boldsymbol{y}_{dr}^{(a)}$. En le concaténant au vecteur des données dans l'échantillon $\boldsymbol{y}_{ds}$, nous construisons le vecteur de recensement $\boldsymbol{y}_d^{(a)} = (\boldsymbol{y}_{ds}', (\boldsymbol{y}_{dr}^{(a)})')'$. Maintenant, en utilisant $\boldsymbol{y}_d^{(a)}$, nous

calculons l'indicateur considéré $\delta_d^{(a)} = \delta_d(\boldsymbol{y}_d^{(a)})$, et on répète l'opération pour $a = 1, \dots, A$. Finalement, l'estimateur HB de δ_d est l'espérance a posteriori, qui est approximée comme suit:

$$\hat{\delta}_d^{HB} = E_{\boldsymbol{y}_{dr}}(\delta_d|\boldsymbol{y}_s) \approx \frac{1}{A}\sum_{a=1}^{A} \delta_d^{(a)}. \tag{57}$$

Puisqu'il n'y pas d'observations dans l'échantillon pour les domaines non échantillonnés ($n_d = 0$), nous avons $\boldsymbol{y}_{dr}^{(a)} = \boldsymbol{y}_d^{(a)}$, et de ce fait nous générons le vecteur complet de recensement $\boldsymbol{y}_d^{(a)} = (Y_{d1}^{(a)}, \dots, Y_{dN_d}^{(a)})'$ à partir de la distribution (55).

Comme mesure de l'erreur d'estimation de l'estimateur HB, $\hat{\delta}_d^{HB}$, la variance a posteriri approchée est obtenue de façon similaire,

$$V(\delta_d|\boldsymbol{y}_s) \approx \frac{1}{A}\sum_{a=1}^{A} \left(\delta_d^{(a)} - \hat{\delta}_d^{HB}\right)^2. \tag{58}$$

Dans le cas spécifique de l'indicateur FGT d'ordre α, $\delta_d = F_{\alpha d}$, dans l'itération a de la procédure de Monte Carlo, nous calculons $F_{\alpha d}^{(a)}$ en utilisant $\boldsymbol{y}_d^{(a)}$ pour appliquer (39) et l'estimateur HB est

$$\hat{F}_{\alpha d}^{HB} \approx \frac{1}{A}\sum_{a=1}^{A} F_{\alpha d}^{(a)}. \tag{59}$$

Comme pour les méthodes ELL et EB, si l'on souhaite estimer un indicateur non linéaire, cette méthode nécessite de disposer, en plus des données d'enquête dont sont extraites les valeurs de la variable d'intérêt, d'un recensement ou de fichiers administratifs à partir desquels on peut obtenir les microdonnées des variables auxiliaires. S'il n'est pas possible d'identifier les individus de l'enquête dans le recensement ou les fichiers, un estimateur HB-recensement peut être calculé de manière similaire à celle de l'estimateur EB-recensement. Dans cet estimateur, même s'il y avait des valeurs appartenant à l'échantillon $\boldsymbol{y}_{ds}$, celles-ci seraient ignorées et le vecteur complet du recensement $\boldsymbol{y}_d^{(a)}$ serait généré, en générant chaque valeur $Y_{di}^{(a)}$ de (55) et la procédure serait la même que si le domaine n'était pas échantillonné.

Résumé des caractéristiques des estimateurs HB fondés sur le modèle avec erreurs emboîtées :

Indicateurs cibles : paramètres généraux.

Données nécessaires :

- Microdonnées relatives aux p variables auxiliaires considérées, provenant de la même enquête que celle où la variable d'intérêt est observée.
- Domaine d'intérêt obtenu à partir de la même enquête que celle où la variable d'intérêt est observée.
- Microdonnées relatives aux p variables auxiliaires considérées provenant d'un recensement ou de fichiers administratifs (mesurées de la même manière que dans l'enquête).

Avantages :

- Ils sont fondés sur des données au niveau individuel, qui fournissent des informations plus détaillées que les données au niveau du domaine (il est également possible d'incorporer des

variables au niveau du domaine). En outre, la taille de l'échantillon est généralement beaucoup plus importante (n par rapport à D).

- Tout type d'indicateur peut être estimé, dès qu'il est défini comme une fonction des variables expliquées Y_{di}.
- Ils sont sans biais sous le modèle si les paramètres du modèle sont connus.
- Ils sont optimaux dans la mesure où ils minimisent la variance a posteriori.
- Dans nos études par simulation, ils s'avèrent être pratiquement égaux aux estimateurs EB.
- Une fois le modèle ajusté, ils peuvent être estimés pour n'importe quelle sous-zone ou sous-domaine. Ils peuvent même être estimés au niveau individuel.
- Une fois le modèle ajusté, tous les indicateurs souhaités (qui sont une fonction des Y_{di}) peuvent être estimés en même temps, sans qu'il soit nécessaire d'ajuster un modèle différent pour chaque indicateur.
- Contrairement à la plupart des procédures bayésiennes, la méthode HB proposée ne nécessite pas l'utilisation de méthodes MCMC et n'exige donc pas de contrôler la convergence des chaînes de Markov.
- Les méthodes bootstrap ne sont pas nécessaires pour l'estimation de l'EQM. Par conséquent, le temps de calcul total peut être nettement inférieur à celui de la méthode EB + bootstrap.
- Le calcul d'intervalles de confiance ou de tout autre résumé de la distribution a posteriori est automatique.

Inconvénients :

- Ils sont fondés sur un modèle. Il est donc nécessaire de vérifier que le modèle s'ajuste correctement (par exemple, par le biais des résidus estimés ou par validation croisée (voir Molina, Nandram, & Rao, 2014)).
- Ils ne prennent pas en compte le plan d'échantillonnage. Ils ne sont pas sans biais sous le plan de sondage et peuvent présenter un biais considérable sous un plan informatif.
- Ils peuvent être sérieusement affectés par des valeurs aberrantes isolées ou par l'absence de normalité.
- La méthode HB ne peut pas être directement étendue à des modèles plus complexes sans perdre certains des avantages mentionnés ci-dessus, notamment celui d'éviter l'application des méthodes MCMC.

F. Méthodes fondées sur des modèles linéaires mixtes généralisés

L'accès à certains services d'éducation ou de santé, ou la disponibilité de certaines commodités de logement, sont généralement mesurés dans une zone particulière en termes de proportion de personnes dans cette zone qui peuvent, ou non, avoir accès au service ou à la commodité en question. Les modèles mixtes linéaires considérés jusqu'à présent ne fournissent pas de prédictions dans l'espace naturel [0,1] où se trouvent ces proportions. Les modèles linéaires mixtes généralisés (GLMM[9]) sont généralement utilisés pour obtenir des prédictions dans cet espace. Si $Y_{di} \in \{0,1\}$ est la variable binaire qui mesure l'absence ou la présence du service ou de la commodité en question, le modèle d'estimation

[9] Generalised linear mixed models.

le plus habituel pour les petits domaines est le modèle GLMM avec effets aléatoires dans les domaines, donné par

$$Y_{di}|v_d \sim \text{Bern}(p_{di}), g(p_{di}) = \boldsymbol{x}_{di}{}'\boldsymbol{\alpha} + v_d, v_d \overset{iid}{\sim} N(0, \sigma_v^2), i = 1, \dots, N_d, d = 1, \dots, D, \tag{60}$$

où v_d est l'effet du domaine d, $\boldsymbol{\alpha}$ est le vecteur des coefficients de la régression et $g: (0,1) \to R$ est la fonction de liaison (bijective, à dérivée continue). En particulier, la liaison logistique, donnée par $g(p) = \log(p/(1-p))$ est probablement la plus largement utilisée en pratique.

Comme on l'a vu plus haut, le meilleur prédicteur sous le modèle (qui minimise l'EQM sous le modèle) du ratio $P_d = \bar{Y}_d$, a pour expression

$$\tilde{P}_d^B(\boldsymbol{\theta}) = E(P_d|\boldsymbol{y}_{ds}; \boldsymbol{\theta}) = \frac{1}{N_d}\left\{\sum_{i \in s_d} Y_{di} + \sum_{i \in r_d} E(Y_{di}|\boldsymbol{y}_{ds}; \boldsymbol{\theta})\right\}. \tag{61}$$

La distribution de $Y_{di}|\boldsymbol{y}_{ds}$ dépend du vecteur $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \sigma_v^2)'$ des paramètres du modèle. En pratique, nous obtenons le prédicteur EB en remplaçant $\boldsymbol{\theta}$ par un estimateur convergent $\widehat{\boldsymbol{\theta}}$ dans la formule du prédicteur optimal, i.e., $\hat{P}_d^{EB} = \tilde{P}_d^B(\widehat{\boldsymbol{\theta}})$.

L'estimateur $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\alpha}}', \hat{\sigma}_v^2)$ de $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \sigma_v^2)$ s'obtient en ajustant le modèle GLMM donné en (60) aux données de l'échantillon $\boldsymbol{y}_s = (\boldsymbol{y}_{1s}', \dots, \boldsymbol{y}_{Ds}')'$. Si l'on veut ajuster le modèle en utilisant la méthode du maximum de vraisemblance, on a besoin de maximiser la vraisemblance donnée par $f(\boldsymbol{y}_s) = \int_{R^D} f(\boldsymbol{y}_s|\boldsymbol{v}) f(\boldsymbol{v}) d\boldsymbol{v}$, où $\boldsymbol{v} = (v_1, \dots, v_D)'$. Sous le modèle GLMM mentionné ci-dessus, une telle vraisemblance n'a pas de forme explicite. Pour assurer cet ajustement, il est donc nécessaire d'utiliser des approximations de l'intégrale (par exemple, numériques) en même temps que des techniques de maximisation numérique. Une fois que le modèle a été ajusté, on doit calculer les espérances $E(Y_{di}|\boldsymbol{y}_{ds}; \widehat{\boldsymbol{\theta}})$ qui définissent le prédicteur EB. Un moyen d'approximer cette espérance serait d'utiliser le théorème de Bayes et le fait que, sachant v_d, les variables $\{Y_{di}; i = 1, \dots, N_d\}$ sont mutuellement indépendantes. Dans ce cas, une telle espérance peut s'écrire comme suit :

$$E(Y_{di}|\boldsymbol{y}_{ds}; \widehat{\boldsymbol{\theta}}) = \frac{E\{h(\boldsymbol{x}_{di}{}'\boldsymbol{\alpha} + v_d) f(\boldsymbol{y}_{ds}|v_d); \widehat{\boldsymbol{\theta}}\}}{E\{f(\boldsymbol{y}_{ds}|v_d); \widehat{\boldsymbol{\theta}}\}}, \quad i \in r_d, \tag{62}$$

où $h = g^{-1}$ est l'inverse de la fonction de liaison et

$$\begin{aligned} f(\boldsymbol{y}_{ds}|v_d) &= \prod_{i \in s_d} p_{di}^{Y_{di}} (1 - p_{di})^{(1-Y_{di})} \\ &= \prod_{i \in s_d} h(\boldsymbol{x}_{di}{}'\boldsymbol{\alpha} + v_d)^{Y_{di}} \{1 - h(\boldsymbol{x}_{di}{}'\boldsymbol{\alpha} + v_d)\}^{(1-Y_{di})}. \end{aligned} \tag{63}$$

Pour la liaison logistique, la fonction inverse est :

$h(\boldsymbol{x}_{di}{}'\boldsymbol{\alpha} + v_d) = \exp(\boldsymbol{x}_{di}{}'\boldsymbol{\alpha} + v_d)/\{1 + \exp(\boldsymbol{x}_{di}{}'\boldsymbol{\alpha} + v_d)\}$. En utilisant (63), on peut approximer les deux espérances qui apparaissent en (62) au moyen d'une simulation de Monte Carlo, en générant $v_d^{(r)} \sim N(0, \hat{\sigma}_v^2)$, $r = 1, \dots, R$, puis en calculant

$$E(Y_{di}|\boldsymbol{y}_{ds}; \widehat{\boldsymbol{\theta}}) \approx \frac{R^{-1} \sum_{r=1}^{R} h(\boldsymbol{x}_{di}{}'\widehat{\boldsymbol{\alpha}} + v_d^{(r)}) \hat{f}(\boldsymbol{y}_{ds}|v_d^{(r)})}{R^{-1} \sum_{r=1}^{R} \hat{f}(\boldsymbol{y}_{ds}|v_d^{(r)})}, \quad i \in r_d, \tag{64}$$

où $\hat{f}$ est la densité conditionnelle $f(\boldsymbol{y}_{ds}|v_d)$, où $\boldsymbol{\alpha}$ est remplacé par $\widehat{\boldsymbol{\alpha}}$.

Le meilleur prédicteur (61) a une EQM minimale et il est sans biais sous le modèle (60). Cependant, l'ajustement du modèle GLMM et le calcul de l'approximation de Monte Carlo de $\hat{P}_d^{EB}$,

comme décrit ci-dessus, nécessitent un temps de calcul important. L'estimation de l'EQM des prédicteurs EB à l'aide d'une procédure de rééchantillonnage augmente le temps de calcul, ce qui la rend peu commode pour les très grandes populations. En outre, lorsque l'on estime les paramètres du modèle $\boldsymbol{\theta}$ et qu'on les remplace par leurs estimateurs afin d'obtenir la version empirique du meilleur prédicteur (EB), on perd l'absence de biais.

Il existe des estimateurs simples qui, bien que n'étant pas optimaux, sont très similaires aux estimateurs optimaux sous certaines conditions et peuvent être obtenus directement en sortie des logiciels usuels d'ajustement des modèles GLMM. Lors de l'estimation d'un ratio, si $\hat{\boldsymbol{\alpha}}$ et $\hat{v}_d$ sont les estimateurs de $\boldsymbol{\alpha}$ et v_d fournis par le logiciel, un estimateur *plug-in* peut être calculé en prédisant simplement les valeurs hors échantillon à l'aide du modèle, c'est-à-dire en supposant

$$\hat{P}_d^{PI} = \frac{1}{N_d}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} \hat{p}_{di}\right), \tag{65}$$

où $\hat{p}_{di} = h(\boldsymbol{x}_{di}'\hat{\boldsymbol{\alpha}} + \hat{v}_d)$ est la valeur prédite de l'observation hors échantillon Y_{di}, $i \in r_d$.

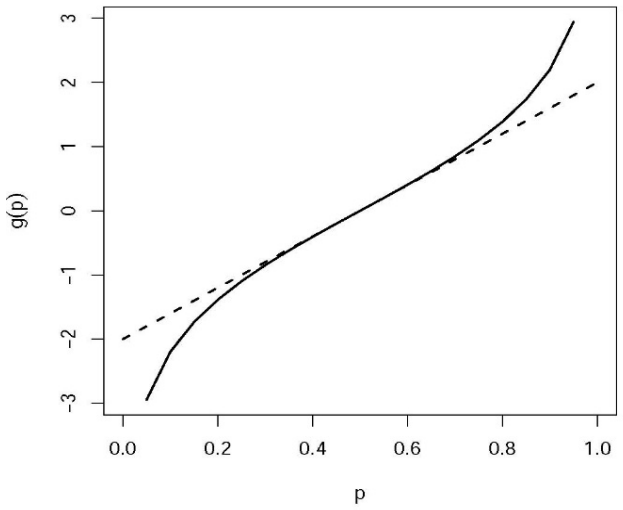
Lorsque $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \sigma_v^2)$ est connu, l'estimateur *plug-in*, $\hat{P}_d^{PI}$, ne peut pas avoir une EQM inférieure à celle du meilleur prédicteur $\hat{P}_d^B$. En effet, contrairement au meilleur prédicteur, l'estimateur *plug-in* n'est pas sans biais à moins que la fonction de liaison soit linéaire. Cependant, l'estimateur *plug-in* est beaucoup plus facile à calculer. Les deux estimateurs correspondent lorsque la fonction de liaison g(·) est linéaire. Dans le cas du lien logistique $g(p) = \log(p/(1-p))$, celle-ci est approximativement linéaire pour $p \in (0.2, 0.8)$ comme le montre la figure 10. Cette linéarité approximative de $g(p)$ pour des valeurs centrales de p nous amène à penser que l'estimateur *plug-in* (65) fondé sur le modèle avec lien logistique devrait être très similaire au prédicteur EB, $\hat{P}_d^{EB}$, en termes d'EQM, au moins pour des valeurs non extrêmes de p. En outre, cette linéarité approximative pour les valeurs centrales de p fait que les estimateurs EB et *plug-in* du ratio $P_d = \bar{Y}_d$, ressemblent à l'estimateur EBLUP, $\hat{P}_d^{EBLUP} = \hat{\bar{Y}}_d^{EBLUP}$, fondé sur le modèle avec erreurs emboîtées décrit au chapitre V. Cela signifie que, pour estimer les proportions d'individus ayant des caractéristiques ni trop peu ni extrêmement fréquentes, il est également logique d'utiliser l'EBLUP.

Les méthodes EB et *plug-in*, fondées sur des modèles non linéaires tels que le GLMM donné dans (60), même pour estimer les valeurs moyennes $\bar{Y}_d$, doivent disposer des valeurs des variables auxiliaires pour tous les individus (microdonnées), obtenues à partir d'un recensement ou de fichiers administratifs. Ceci est nécessaire pour calculer l'espérance $E(Y_{di}|\boldsymbol{y}_{ds}; \boldsymbol{\theta})$ dans le cas du prédicteur EB, ou pour prédire la probabilité $\hat{p}_{di}$ dans le cas de l'estimateur *plug-in*. Cependant, en plus des données d'enquête, l'estimateur EBLUP de $\bar{Y}_d$ ne nécessite que les moyennes de ces variables dans la population à l'intérieur des domaines. Ces données agrégées sont généralement disponibles sans restriction de confidentialité.

En principe, le modèle GLMM donné dans (60) pourrait être utilisé pour estimer le taux de pauvreté (indicateur FGT d'ordre $\alpha = 0$), F_{0d}. Pour l'écart de pauvreté (indicateur FGT d'ordre $\alpha = 1$), F_{1d}, il ne serait pas judicieux de l'utiliser car il ne s'agit pas de ratios, puisque les valeurs individuelles $F_{1,di}$ ne sont pas des variables binaires. Dans le cas du taux de pauvreté, en prenant $Y_{di} = I(E_{di} < z)$ comme variable expliquée binaire, nous obtenons $P_d = F_{\alpha d}$. Le meilleur prédicteur résultant repose sur l'expression (47) de la section V.D, mais l'espérance apparaissant dans le deuxième terme serait prise par rapport à la distribution conditionnelle sous le modèle (60) et devrait être approximée numériquement ; par exemple, comme dans (64) puisque, dans ce cas, les distributions conditionnelles $Y_{di}|\boldsymbol{y}_{ds}$ n'ont pas une forme connue. Comme cela a été dit, l'estimateur *plug-in* (65) aurait un coût de calcul plus faible. Encore une fois, si les unités d'enquête ne peuvent être identifiées dans le recensement ou les fichiers, un estimateur EB-recensement peut être utilisé pour remplacer les

observations de l'échantillon dans le prédicteur (61) Y_{di}, $i \in s_d$, avec les prédictions obtenues comme dans (62), ou en utilisant $\hat{p}_{di}$ comme prédiction dans le cas de l'estimateur *plug-in* - recensement.

Figure 10
Liaison logistique



Source : D'après l'auteur.

L'EQM du prédicteur correspondant (que ce soit EB ou *plug-in*) peut être estimée en utilisant une procédure bootstrap comme suit (voir González-Manteiga et al., 2007) :

- Ajuster le modèle GLMM donné en (60) aux données de l'échantillon s, pour obtenir les estimateurs $\hat{\sigma}_v^2$ et $\hat{\boldsymbol{\alpha}}$ des paramètres du modèle.
- Générer les effets aléatoires bootstrap
$$v_d^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_v^2), \quad d = 1, \dots, D.$$
- Générer un recensement bootstrap $\boldsymbol{y}_d^{*(b)} = (Y_{d1}, \dots, Y_{dN_d})'$, comme suit :
$$Y_{di}^{*(b)} \overset{ind}{\sim} \text{Bern}(p_{di}^{*(b)}), p_{di}^{*(b)} = h(\boldsymbol{x}_{di}'\hat{\boldsymbol{\alpha}} + v_d^{*(b)}), i = 1, \dots, N_d, d = 1, \dots, D, \qquad (66)$$
et calculer les vraies valeurs des indicateurs $P_d^{*(b)} = \bar{Y}_d^{*(b)}$, $d = 1, \dots, D$.
- Pour chaque domaine $d = 1, \dots, D$, extraire les éléments de ce domaine appartenant à l'échantillon, à partir du recensement bootstrap $\boldsymbol{y}_d^{*(b)}$, Y_{di}, $i \in s_d^{*(b)}$, en construisant le vecteur $\boldsymbol{y}_{ds}^{*(b)}$. Soit $\boldsymbol{y}_s^{*(b)} = ((\boldsymbol{y}_{1s}^{*(b)})', \dots, (\boldsymbol{y}_{Ds}^{*(b)})')'$ le vecteur des valeurs dans l'échantillon pour tous les domaines, où $s = s_1 \cup \cdots \cup s_D$ est l'échantillon d'origine.
- Ajuster le modèle (60) aux données bootstrap $\boldsymbol{y}_s^{*(b)}$ et calculer les prédicteurs bootstrap $\hat{P}_d^{EB*(b)}$, $d = 1, \dots, D$.
- Répéter les étapes 2) - 5), pour $b = 1, \dots, B$. L'estimateur bootstrap de l'EQM du prédicteur $\hat{P}_d^{EB}$ s'exprime comme
$$mse_B(\hat{P}_d^{EB}) = B^{-1} \sum_{b=1}^{B} (\hat{P}_d^{EB*(b)} - P_d^{*(b)})^2.$$

Résumé des caractéristiques du prédicteur fondé sur GLMM/*plug-in* comparées aux méthodes applicables à l'estimation d'une moyenne

Indicateurs cibles : Proportions ou totaux d'une variable binaire (par exemple, non disponibilité d'un certain produit ou service).

Données nécessaires :

- Microdonnées relatives aux p variables auxiliaires considérées, provenant de la même enquête que celle où la variable d'intérêt est observée.
- Domaine d'intérêt obtenu à partir de la même enquête que celle où la variable d'intérêt est observée.
- Microdonnées relatives aux p variables auxiliaires considérées provenant d'un recensement ou de fichiers administratifs (mesurées de la même manière que dans l'enquête).

Avantages:

- Le nombre d'observations utilisées pour ajuster le modèle est la taille totale de l'échantillon n, beaucoup plus grande que le nombre de domaines dans les modèles FH. Les paramètres du modèle sont donc estimés très efficacement et les améliorations de l'efficacité par rapport aux estimateurs directs seront plus importantes qu'avec les modèles FH.
- Le modèle de régression considéré incorpore l'hétérogénéité inexpliquée entre les domaines.
- Contrairement au modèle FH, il n'est pas nécessaire de connaître la variance.
- L'estimateur de l'EQM obtenu sous le modèle (par exemple, au moyen de procédures bootstrap) est un estimateur stable sous le modèle et il est sans biais sous le modèle lorsqu'il est moyenné sur de nombreux domaines.
- Les estimations peuvent être désagrégées pour tout sous-domaine ou sous-zone souhaités au sein des domaines, et même au niveau individuel.
- On peut faire des estimations sur des domaines non échantillonnés.

Inconvénients :

- Les prédicteurs considérés sont fondés sur un modèle, et il est donc nécessaire d'analyser ce modèle (par exemple, au moyen des résidus).
- Ils ne prennent pas en compte le plan d'échantillonnage. Par conséquent, ils ne sont pas sans biais sous le plan de sondage et ils sont mieux adaptés à un échantillonnage aléatoire simple. Ils seront affectés par des plans de sondage informatifs.
- Les microdonnées sont généralement obtenues à partir d'un recensement ou de fichiers administratifs, et il existe souvent des problèmes de confidentialité qui limitent l'utilisation de ce type de données.
- L'estimation de l'EQM obtenue sous le modèle (par exemple, au moyen de procédures bootstrap) est correcte sous le modèle considéré et n'est pas sans biais sous le plan pour un domaine donné.
- Le prédicteur EB (contrairement à l'estimateur plug-in) a un coût de calcul élevé.

- L'EQM du prédicteur EB que l'on obtient (par exemple, au moyen d'une procédure bootstrap) a un coût de calcul excessivement élevé et peut être peu adaptée pour de très grandes populations. Ce coût est nettement inférieur pour le prédicteur plug-in.
- Les estimateurs nécessitent un réajustement pour vérifier la propriété d'additivité, de façon à ce que la somme des totaux estimés dans les domaines au sein d'une plus grande région corresponde à l'estimateur direct pour cette zone.

VI. Application : estimation du revenu moyen et des taux de pauvreté à Montevideo

Dans ce chapitre, nous allons utiliser certaines des techniques décrites ci-dessus pour estimer les revenus moyens et le taux de pauvreté non extrême pour les secteurs de recensement et pour les deux sexes à Montevideo, en Uruguay. Pour ce faire, nous utiliserons les données de l'enquête en continu auprès des ménages (*Encuesta Continua de Hogares* ou *ECH*) et du recensement de la population, tous deux de 2011. Cette application n'a qu'un but illustratif et peut probablement être améliorée en effectuant une recherche plus approfondie d'informations auxiliaires. Par conséquent, les résultats obtenus dans cette application ne doivent pas être considérés comme des estimations définitives.

Puisque les paramètres des modèles à considérer peuvent dépendre du sexe, pour chaque type d'estimateur, nous ajusterons des modèles distincts pour chaque sexe. Plus précisément, nous calculerons des estimations directes en utilisant les microdonnées de l'ECH pour chaque secteur et chaque sexe, des estimations FH fondées sur le modèle de base au niveau domaine (21), en utilisant certains totaux de population issus du recensement comme information auxiliaire pour chaque secteur et chaque sexe et, enfin, des estimations EB-recensement fondées sur le modèle de base au niveau individuel (38) pour le logarithme du revenu, en utilisant les microdonnées du recensement pour certaines variables également mesurées dans l'ECH. Notons que, même si l'on estimait seulement le revenu moyen, qui est un paramètre s'exprimant linéairement en fonction du revenu des individus de la zone, lorsqu'on effectue une transformation non linéaire (logarithmique) de la variable expliquée dans le modèle avec erreurs emboîtées (38), le paramètre cible, exprimé en fonction des valeurs de la variable expliquée du modèle, sera une fonction non linéaire des valeurs de cette variable. Ainsi, dans ce cas, l'EBLUP n'a pas de sens puisqu'il s'agit d'un estimateur linéaire en les valeurs de la variable expliquée par

le modèle dans l'échantillon et nous devons recourir à la méthode EB. De plus, comme les individus de l'ECH ne sont pas identifiés dans le recensement, nous considérons l'estimateur EB-recensement. En plus des estimateurs ponctuels, nous obtiendrons des estimations des EQM de chaque estimateur. Les calculs ont été effectués à l'aide des *packages R sae* (Molina et Marhuenda, 2015) et *lme4* (Bates et al. 2015).

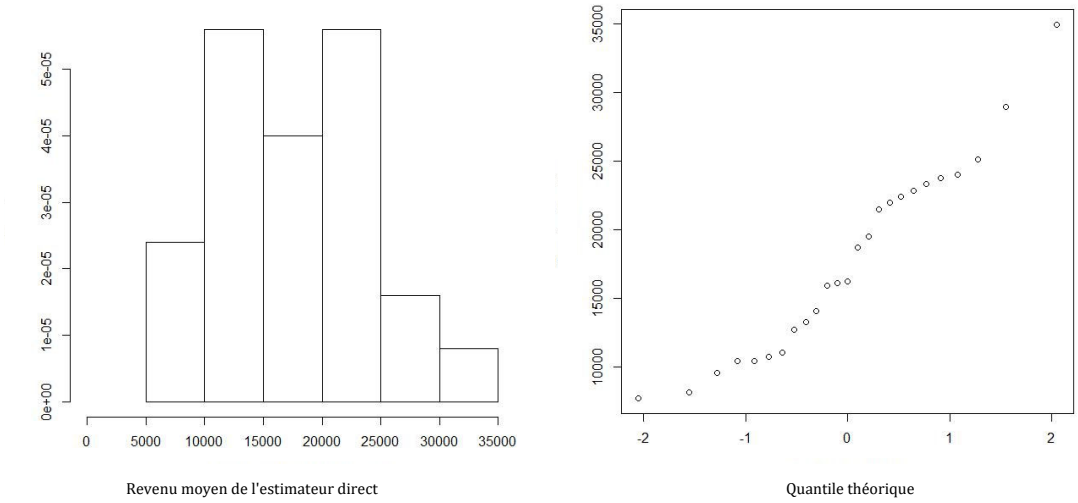
Les tailles de population selon le recensement complet de 2011 à Montevideo (pour les habitants en logement privé) sont N=656 162 pour les femmes et N=566 698 pour les hommes. La taille des échantillons ECH, après élimination des données manquantes, est de n=26 233 pour les femmes et n=22 464 pour les hommes. Pour les secteurs D=25 qui apparaissent dans le recensement, la taille des échantillons varie entre 56 et 3482 pour les femmes et entre 65 et 2820 pour les hommes. Bien qu'il ne s'agisse pas de tailles d'échantillon excessivement petites, nous verrons que les techniques d'estimation sur petits domaines peuvent toujours fournir des estimations plus précises, en mesurant cette précision en termes d'erreur quadratique moyenne. Il faut également garder à l'esprit que, selon les données disponibles, le taux de pauvreté à Montevideo est relativement faible et que, pour estimer ces chiffres avec précision à l'aide d'estimateurs directs, les tailles d'échantillon nécessaires par secteur et par sexe doivent être plus importantes que pour estimer des proportions proches de 0,5 ou des moyennes de variables continues, comme le revenu moyen. En fait, même si la taille de l'échantillon n'est pas excessivement petite, l'estimateur direct peut prendre une valeur nulle du fait qu'aucun individu n'est observé avec des revenus inférieurs au seuil de pauvreté. Le seuil de pauvreté non extrême pour les zones urbaines en 2011 est de 3 182 pesos uruguayens.

Aussi bien pour le revenu moyen $\bar{E}_d = N_d^{-1} \sum_{i=1}^{N_d} E_{di}$ que pour le taux de pauvreté $F_{0d} = N_d^{-1} \sum_{i=1}^{N_d} I\,(E_{di} < z)$ pour chaque secteur de recensement et chaque sexe, les estimateurs directs correspondants , $\hat{\bar{E}}_d^{DIR}$ et $\hat{F}_{0d}^{DIR}$, et leurs variances d'échantillon estimées $\widehat{\text{var}}_\pi(\hat{\bar{E}}_d^{DIR})$ et $\widehat{\text{var}}_\pi(\hat{F}_{0d}^{DIR})$ sont obtenus en utilisant les microdonnées ECH dans les formules 4) - 6). Ceci est fourni par la fonction direct() du package *sae*, en introduisant les poids d'échantillonnage de l'ECH. Pour les tailles de population des secteurs de recensement, N_d, nous utilisons les tailles obtenues à partir du recensement.

Les estimateurs FH et leurs erreurs quadratiques moyennes estimées sont obtenus à partir du modèle (21) pour $\delta_d = \bar{E}_d$ ou $\delta_d = F_{0d}$. Dans le cas du revenu moyen, $\delta_d = \bar{E}_d$, pour les deux sexes, nous considérons comme variables auxiliaires agrégées au niveau du secteur de recensement (composantes de $\boldsymbol{x}_d$ dans le modèle), les proportions issues du recensement des individus alphabétisés, des individus au chômage (mais non retraités), l'âge moyen et le nombre moyen d'années passées dans le système éducatif. Pour le taux de pauvreté, $\delta_d = F_{0d}$, seules les proportions de personnes alphabétisées et de personnes au chômage sont significatives. Les estimateurs FH peuvent être obtenus en utilisant la fonction eblupFH() du package *sae* qui implémente la formule donnée dans (24). Comme vecteur des variables expliquées du modèle, on établit le vecteur des estimations directes précédemment obtenues $\hat{\bar{E}}_d^{DIR}$ ou $\hat{F}_{0d}^{DIR}$ selon le cas et, pour les variances ψ_d, les estimations des variances de l'échantillon $\widehat{\text{var}}_\pi(\hat{\bar{E}}_d^{DIR})$ ou $\widehat{\text{var}}_\pi(\hat{F}_{0d}^{DIR})$. Les EQM estimées, $\text{mse}_{PR}(\hat{\delta}_d^{FH})$, sont obtenues en utilisant les formules analytiques de la section V.A, pour l'ajustement par la méthode REML, et dans R, elles sont obtenues en utilisant la fonction mseFH() du package ci-dessus.

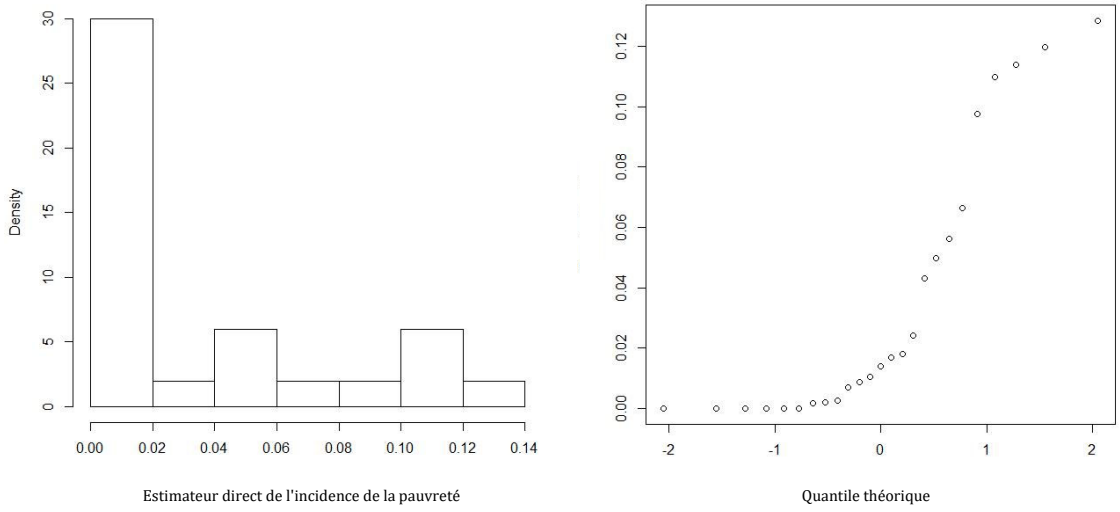
Les estimateurs FH du taux de pauvreté peuvent prendre la valeur zéro (tout comme les estimateurs directs) dans les domaines où il n'y a pas d'individus dont le revenu est inférieur au seuil de pauvreté. De plus, les EQM estimées au moyen de la formule analytique ci-dessus prennent également la valeur zéro. Dans ces cas, nous considérons que de telles estimations FH ne sont pas fiables et, à la place, nous calculons des estimateurs synthétiques $\hat{\delta}_d^{FH} = \boldsymbol{x}_d{}'\widehat{\boldsymbol{\beta}}$. Leurs EQM sont obtenues en utilisant la formule (6.2.14) de Rao et Molina (2015), en remplaçant l'estimateur REML de la variance des effets de domaine.

Figure 11
Histogramme (à gauche) et diagramme de normalité q-q (à droite) des estimateurs directs du revenu moyen pour les $D = 25$ secteurs de recensement de Montevideo, pour les femmes
(Pesos uruguayens, année 2011)



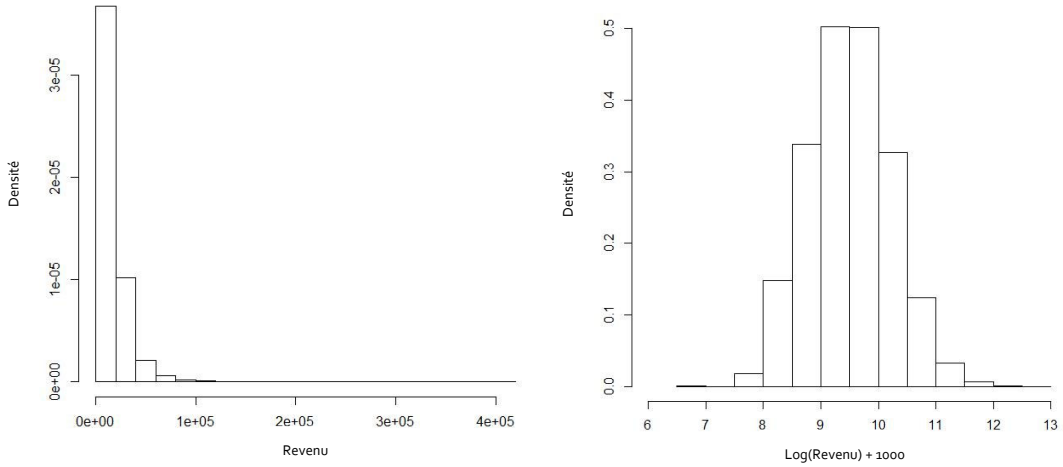
Source : D'après l'auteur.

Figure 12
Histogramme (à gauche) et diagramme de normalité q-q (à droite) des estimateurs directs du taux de pauvreté non extrême pour les $D = 25$ secteurs de recensement de Montevideo, pour les femmes
(En proportions)



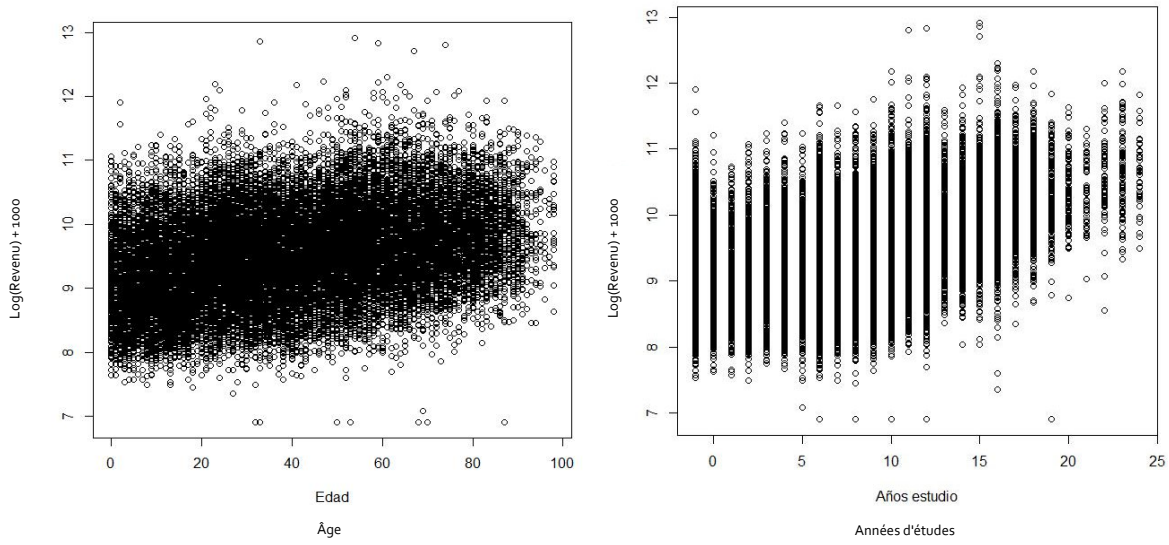
Source : D'après l'auteur.

Figure 13
Histogramme du revenu non transformé (à gauche) et transformé en logarithme (revenu + 1000) (à droite) pour les femmes
(Pesos uruguayens, année 2011)



Source : D'après l'auteur.

Figure 14
Revenu transformé en comparaison avec l'âge (à gauche) et en fonction du nombre d'années d'études (à droite), pour les femmes
(Pesos uruguayens - transformation logarithmique -, année 2011)



Source : D'après l'auteur.

Bien que l'estimateur EBLUP basé sur le modèle de Fay-Herriot n'exige pas la normalité, l'approximation analytique de l'EQM obtenue de cette manière exige la normalité. Comme nous pouvons le voir dans l'histogramme et le graphique q-q de normalité pour les femmes (figure 11), la distribution des estimateurs directs du revenu moyen pour les secteurs de recensement D=25 n'est pas exactement une distribution normale, mais elle n'en est pas trop éloignée non plus, compte tenu du fait

que le nombre d'observations utilisé pour construire l'histogramme (D=25) est faible. Pour les hommes, les graphiques sont similaires. Ce n'est pas le cas pour les estimateurs directs du taux de pauvreté non extrême (voir figure 12). Il faut donc garder à l'esprit que les EQM estimées de ces taux de pauvreté peuvent ne pas correspondre à la réalité.

Enfin, nous obtenons les estimateurs EB-recensement fondés sur le modèle au niveau individuel (38), en utilisant comme variable expliquée $\log(\text{income} + 1000)$ où l'ajout de la constante 1000 au revenu a été déterminé de manière à ce que l'histogramme des résidus du modèle ajusté soit approximativement symétrique (voir l'histogramme du revenu original et du revenu transformé de cette manière dans la Figure 13). Comme variables auxiliaires au niveau individuel, soit les $\boldsymbol{x}_{di}$,, nous considérons les indicateurs du statut d'activité, l'âge et le nombre d'années d'études. La figure 14 pour les femmes montre une relation croissante approximativement linéaire entre les revenus transformés et l'âge ou le nombre d'années d'études. Le graphique pour les hommes est similaire. Comme la transformation des gains est monotone, cette relation indique que les gains augmentent avec l'âge ou le nombre d'années d'études.

Les estimateurs EB du recensement du taux de pauvreté $F_{\alpha d}$ fondés sur le modèle avec erreurs emboîtées pour le revenu transformé sont calculés à l'aide des formules (53) et (48), en remplaçant $\boldsymbol{\theta}$ par l'estimateur $\widehat{\boldsymbol{\theta}}$; dans ce cas, nous avons utilisé l'estimateur REML. Bien que, comme nous l'avons vu dans l'exemple 7, la fonction ebBHF() du package *sae* fournisse les estimateurs EB mais pas les estimateurs EB du recensement, si les fractions d'échantillon dans les domaines sont petites, nous pouvons utiliser la même fonction pour obtenir des valeurs approximatives des estimateurs EB du recensement, en prenant pour argument Xnonsample de cette fonction (matrice des valeurs des variables auxiliaires pour la partie hors échantillon de la population) égale à la matrice des microdonnées du recensement relatives à ces variables pour tous les individus dans les secteurs considérés. Dans ce cas, on peut constater que les estimations EB-recensement et celles obtenues de cette manière présentent des différences vraiment faibles.

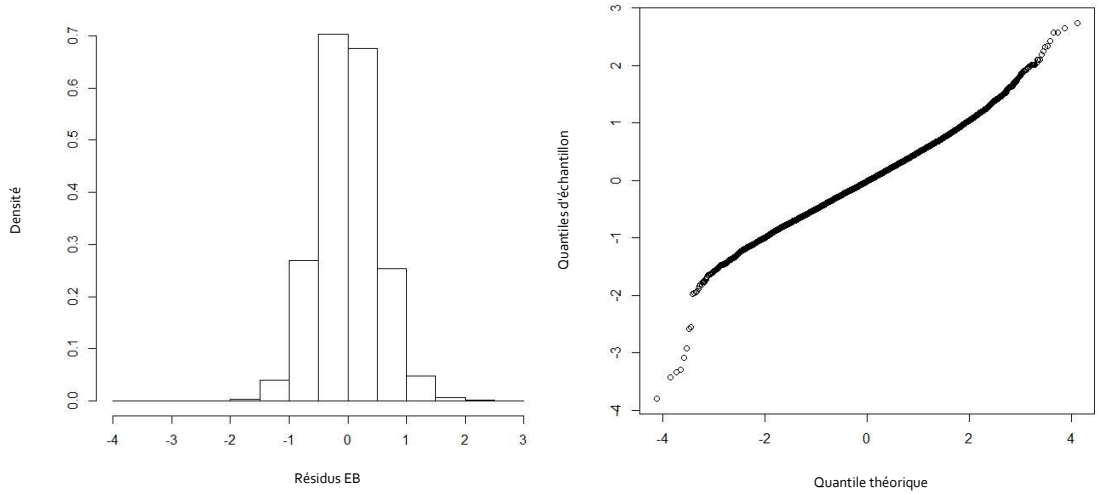
Le même modèle ajusté est utilisé pour obtenir les estimateurs EB du recensement des gains moyens. Les estimateurs EB du recensement des gains moyens $\delta_d = \bar{E}_d$ fondés sur ce modèle sont obtenus de manière similaire. Plus précisément, ils sont obtenus comme suit : $\hat{\bar{E}}_d^{CEB} = N_d^{-1} \sum_{i=1}^{N_d} \hat{E}_{di}^{CEB}$ où, en tenant compte du fait que le revenu E_{di} s'écrit en fonction des variables expliquées du modèle Y_{di} comme suit : $E_{di} = \exp(Y_{di}) + 1000$, alors : $\hat{E}_{di}^{CEB} = E[\exp(Y_{di})|\boldsymbol{y}_s; \widehat{\boldsymbol{\theta}}] + 1000$. Cette espérance peut être obtenue en utilisant l'approximation de Monte Carlo (50) implémentée dans la fonction ebBHF(), ou au moyen de la formule analytique donnée dans Molina et Martín (2018). Dans le cas présent, cette formule analytique a été utilisée car elle ne génère aucun coût de calcul.

Les EQM estimées des estimateurs EB du recensement sont obtenues au moyen d'une légère modification de la procédure bootstrap décrite au chapitre V (initialement conçue pour les estimateurs EB), en utilisant $B = 500$ réplications bootstrap. La différence entre les estimateurs EB et EB-recensement réside dans le fait que les unités ECH ne peuvent pas être identifiées dans le recensement. Par conséquent, dans chaque réplication bootstrap, nous ne pouvons pas générer les vecteurs de recensement $\boldsymbol{y}_d^{*(b)}$, $d = 1, \dots, D$, et en extraire la partie relative à l'échantillon $\boldsymbol{y}_s^{*(b)}$. Dans le cas des estimateurs EB-recensement, nous générons les recensements bootstrap $\boldsymbol{y}_d^{*(b)}$, $d = 1, \dots, D$, en utilisant les valeurs des variables auxiliaires du recensement et, d'autre part, nous générons le vecteur bootstrap de l'échantillon $\boldsymbol{y}_s^{*(b)}$ en utilisant les valeurs des mêmes variables auxiliaires, mais tirées de l'ECH. Les véritables valeurs bootstrap des paramètres sont obtenues à partir des recensements bootstrap générés, $\delta_d^{(b)} = \delta_d(\boldsymbol{y}_d^{*(b)})$, $d = 1, \dots, D$.

La méthode EB (ou EB-recensement) et la procédure de bootstrap utilisée reposent sur l'hypothèse de normalité ; il est donc, dans ce cas, crucial de vérifier si cette hypothèse est vérifiée, au

moins approximativement. La figure 15 montre l'histogramme et le graphique de normalité q-q des résidus de l'ajustement du modèle pour les gains transformés des femmes. Bien que les données réelles s'adaptent difficilement à un modèle exact, et que tout test rejettera l'hypothèse nulle de normalité si la taille de l'échantillon est aussi grande que dans ce cas, nous pouvons voir sur ces figures que la distribution n'est pas trop éloignée de la normale. Si l'on souhaite utiliser une distribution qui correspond un peu mieux au revenu, on peut utiliser la méthode EB fondée sur un modèle GB2 multivarié tel que celui proposé par Graf, Marín et Molina (2018).

Figure 15
Histogramme (à gauche) et diagramme de normalité q-q (à droite) des résidus du modèle avec erreurs emboîtées pour le revenu transformé, pour les femmes
(En proportions)



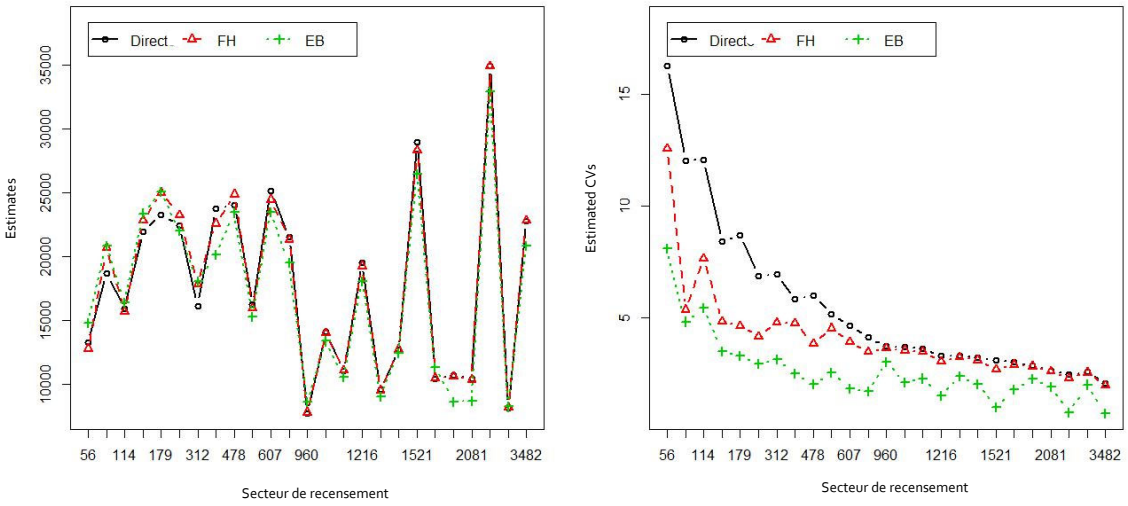
Source: D'après l'auteur.

Les résultats numériques détaillés pour chaque secteur de recensement sont présentés dans les tableaux 1-4 de l'annexe. Nous analysons ces résultats graphiquement et commentons les résultats obtenus pour les différents estimateurs. La figure 16 montre les valeurs obtenues à partir des estimateurs directs, FH et EB-recensement, du revenu moyen (à gauche), et les CV estimés de ces estimateurs (à droite) pour chaque secteur de recensement, pour les femmes. Les secteurs de recensement (axe x) sont classés de la plus petite à la plus grande taille d'échantillon et les tailles d'échantillon ont été indiquées sur l'axe x. Nous pouvons voir que les trois estimateurs prennent des valeurs similaires, bien que les estimateurs direct et FH conduisent pratiquement aux mêmes valeurs dans ce cas. Cela est dû au fait que, lors de l'estimation du revenu moyen, les tailles d'échantillon des secteurs ne sont pas excessivement petites, et que le poids donné par les estimateurs FH aux estimateurs directs correspondants est proche de un. C'est un avantage des estimateurs fondés sur des modèles à effets aléatoires. Cependant, et bien que la taille des échantillons soit modérée, comme nous pouvons le voir dans le graphique de droite, les estimateurs EB du recensement sont clairement plus efficaces que les estimateurs directs et FH pour tous les secteurs de recensement. Cela est dû au fait qu'ils utilisent une plus grande quantité d'informations : les microdonnées du recensement. Pour les hommes (figure 17), nous pouvons obtenir des conclusions similaires.

Pour le taux de pauvreté, les estimations et les erreurs quadratiques moyennes pour les femmes et les hommes sont présentées dans les Figures 18 et 19 respectivement. Dans ce cas, nous faisons figurer les EQM au lieu des CV parce que, dans le cas des ratios, pour une taille d'échantillon fixée, les

CV augmentent à mesure que le ratio diminue ; par conséquent, les CV sont moins significatifs en tant que mesures de l'erreur d'estimation, surtout lorsque les proportions estimées prennent des valeurs très faibles, comme c'est le cas ici. Encore une fois, les valeurs des trois estimateurs sont similaires pour tous les secteurs de recensement, sauf pour ceux dont la taille d'échantillon est la plus petite. En fait, dans ces secteurs, les estimateurs directs pour les femmes prennent une valeur nulle (peu plausible) parce qu'il n'y a aucun individu échantillonné dont le revenu est inférieur au seuil. En fait, les variances estimées des estimateurs directs prennent également la valeur nulle pour ces secteurs. Il faut noter que les variances estimées des estimateurs directs sont également basées sur les quelques observations échantillonnées pour chaque quartier et chaque genre. Si nous considérons que les estimateurs directs ne sont pas fiables, leurs variances estimées ne le sont pas non plus. Pour les domaines dont les estimateurs directs sont égaux à zéro, les estimateurs FH et leurs EQM sont aussi théoriquement nuls. Dans ces cas, comme mentionné ci-dessus, les estimateurs synthétiques obtenus à partir du même modèle ont été utilisés. Nous pouvons observer dans les figures de droite que les EQM des estimateurs directs et FH montrent de grandes fluctuations. Il faut noter que les EQM des estimateurs FH sont particulièrement importantes pour les domaines où des estimateurs synthétiques ont été utilisés. En revanche, les EQM des estimateurs EB augmentent faiblement en fonction de la taille de l'échantillon du secteur de recensement. En plus de prendre des valeurs plus raisonnables, les EQM estimées des estimateurs EB restent inférieures aux EQM des deux autres estimateurs pour la plupart des secteurs de recensement.

Figure 16
Estimations directes, FH et EB-recensement (à gauche) du revenu moyen, et CV des estimateurs (à droite) pour les $D = 25$ secteurs de recensement de Montevideo, pour les femmes
(Pesos uruguayens, année 2011)

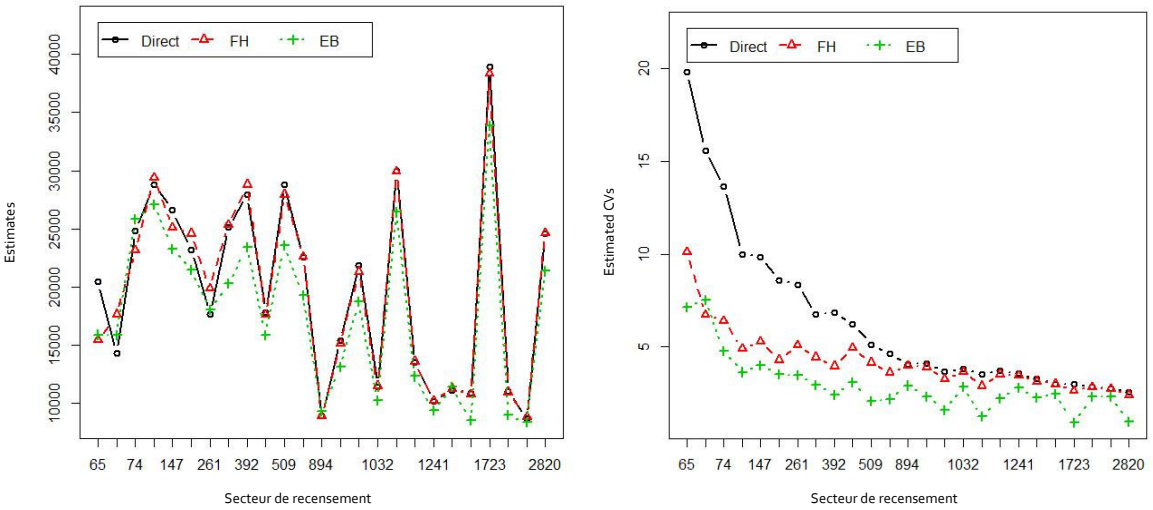


Source : D'après l'auteurr.
Note : Les secteurs de recensement sont ordonnés (axe des x) de la plus petite à la plus grande taille d'échantillon, les tailles d'échantillon étant indiquées sur l'axe.

Il convient de souligner que les estimateurs fondés sur un modèle peuvent même fournir des estimations pour des domaines non échantillonnés, bien que cela ne soit pas recommandé car il n'est pas possible d'analyser la qualité de l'ajustement du modèle pour ces domaines. Et, sur la qualité de l'ajustement, comme indiqué ci-dessus, pour le taux de pauvreté non extrême, l'hypothèse de normalité dans le modèle de Fay-Herriot n'est pas vérifiée. Ceci est dû au fait que la taille des échantillons est faible pour certains des secteurs et que les taux réels de pauvreté semblent assez faibles, avec pour

conséqeunce que les estimateurs directs ont une distribution nettement asymétrique et que le théorème central limite n'est pas vérifié. Bien que la normalité ne soit pas une exigence pour obtenir l'estimateur FH, on doit en faire l'hypothèse pour l'estimation de l'EQM en utilisant les formules analytiques indiquées dans la section V.A et fournies par la fonction *mseFH()* du package *sae*. En fait, un inconvénient supplémentaire des estimateurs obtenus à partir de ce modèle de Fay-Herriot est qu'ils peuvent donner des valeurs négatives ou supérieures à un, ce qui, lorsqu'il s'agit de ratios, n'est pas approprié. Une solution simple consiste à tronquer les estimations à zéro lorsqu'elles sont négatives et à un lorsqu'elles dépassent cette valeur.

Figure 17
Estimations directes, FH et EB-recensement (à gauche) du revenu moyen, et CV des estimateurs (à droite) pour les $D = 25$ secteurs de recensement de Mondevideo, pour les hommes
(Pesos uruguayens, année 2011)



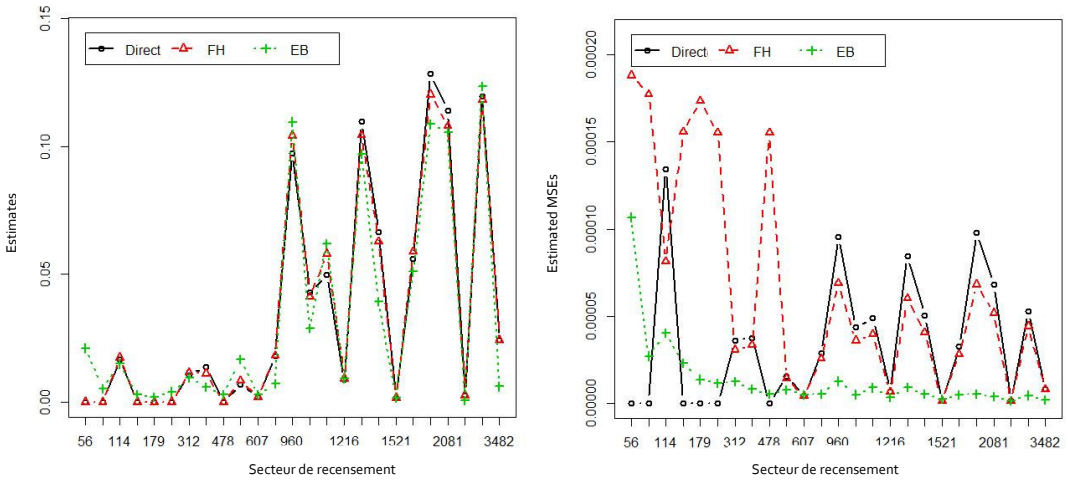
Source : D'après l'auteurr.
Note : Les secteurs de recensement sont ordonnés (axe des x) de la plus petite à la plus grande taille d'échantillon, les tailles d'échantillon étant indiquées sur l'axe.

Une autre possibilité est de considérer le modèle de régression (19) pour une transformation bijective du taux de pauvreté, $g(F_{0d})$, qui transforme les valeurs de l'espace [0,1] en valeurs réelles. Cependant, la même transformation de l'estimateur direct, $g(\hat{F}_{0d}^{DIR})$, ne doit pas nécessairement être un estimateur sans biais de $g(F_{0d})$, et, par conséquent, le modèle (20) n'est pas vérifié pour $g(\hat{F}_{0d}^{DIR})$. Dans ce cas, le modèle FH pour $g\left(\hat{F}_{0d}^{DIR}\right)$ aurait un biais supplémentaire, sauf si l'on considère le modèle (20) pour $\hat{F}_{0d}^{DIR}$ conjointement avec le modèle de régression ci-dessus pour $g(F_{0d})$. Dans ce cas, les deux modèles considérés ne peuvent pas être résumés dans un modèle mixte linéaire tel que celui donné dans (21), c'est-à-dire qu'il s'agit de modèles non appariés. Des estimateurs basés sur des modèles non appariés de ce type ont été obtenus par You et Rao (2002b) sur la base de l'inférence bayésienne.

Comme nous l'avons vu, lorsque nous disposons d'informations auxiliaires au niveau individuel, on peut encore améliorer l'efficacité des estimateurs qui utilisent ces informations. Cependant, deux sources de données relatives à la même année ont été utilisées dans cette application. Les années où un recensement actualisé n'est pas disponible, les estimateurs basés sur des modèles au niveau individuel peuvent fournir des estimations quelque peu biaisées. Dans ces cas, il est donc conseillé de rechercher d'autres sources de données actualisées, telles que celles issues des fichers administratifs. Lorsqu'il n'existe pas de sources de données actualisées au niveau individuel, il est recommandé de s'en tenir à

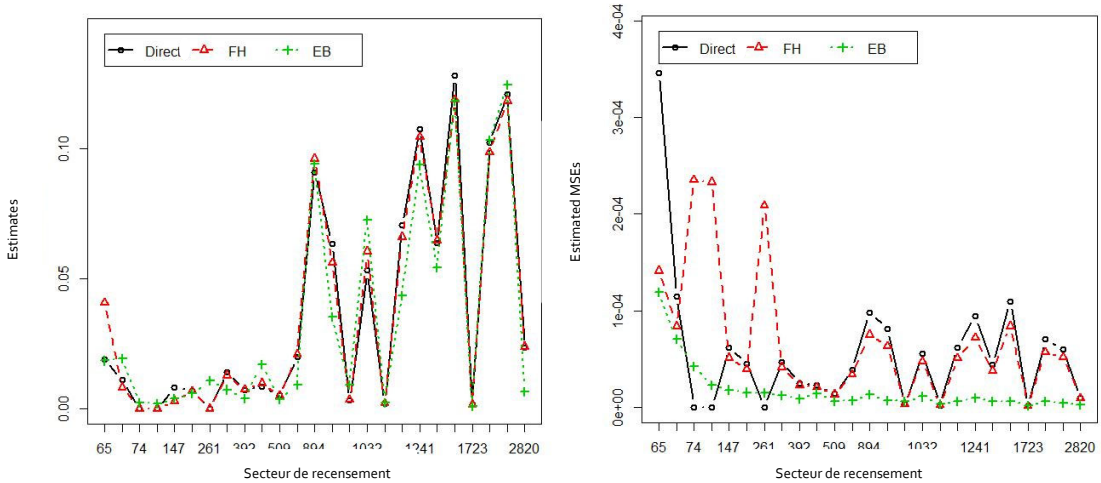
des modèles au niveau du domaine. Dans certains cas, il est possible de trouver des sources de données agrégées à un niveau inférieur à celui de la zone considérée. Dans ce cas, des modèles relatifs aux données agrégées pourraient être utilisés à ce niveau, notamment des modèles à deux niveaux de zone (voir Torabi et Rao (2014)).

Figure 18
Estimations directes, FH et EB-recensement (à gauche) des taux de pauvreté, et EQM des estimateurs (à droite) pour les $D = 25$ secteurs de recensement de Montevideo, pour les femmes



Source : D'après l'auteurr.
Note : Les secteurs de recensement sont ordonnés (axe des x) de la plus petite à la plus grande taille d'échantillon, les tailles d'échantillon étant indiquées sur l'axe.

Figure 19
Estimations directes, FH et EB-recensement (à gauche) des taux de pauvreté, et EQM des estimateurs (à droite) pour les $D = 25$ secteurs de recensement de Montevideo, pour les hommes
(En proportions)



Source : D'après l'auteurr.
Note : Les secteurs de recensement sont ordonnés (axe des x) de la plus petite à la plus grande taille d'échantillon, les tailles d'échantillon étant indiquées sur l'axe.

VII. Conclusions

Ce document a traité du problème de la désagrégation des estimations statistiques sur des domaines ou des sous-groupes de population. Il fournit des recommandations sur les limites de la désagrégation des estimations directes et une description des méthodes indirectes de base, ainsi que de certaines méthodes plus sophistiquées, qui peuvent réduire ces limites. Comme nous l'avons vu tout au long de ce document, les méthodes à utiliser dans chaque application spécifique dépendent principalement de la forme de l'indicateur considéré et du type d'informations auxiliaires disponibles, car il n'existe pas de méthodes universelles pouvant être utilisées pour tout type d'indicateur ou d'information disponible. Ainsi, dans chaque cas, une étude doit être faite des méthodes potentiellement applicables, en fonction des exigences en matière de données et des hypothèses que chaque méthode suppose. Dans les applications qui permettent l'utilisation de diverses méthodes, la précision des estimateurs finaux dépendra de jusqu'à quel degré les variables auxiliaires disponibles peuvent être considérées comme de bons prédicteurs de la variable modélisée dans chaque cas et dans quelle mesure les hypothèses correspondantes sont vérifiées.

Il ne faut pas oublier que, si l'on exige des estimations aussi précises que possible, la mesure des erreurs correspondantes (généralement les erreurs quadratiques moyennes) doit également être estimée aussi précisément que possible ou , à tout le moins, il ne faut pas sous-estimer ces erreurs, afin de ne pas donner une image faussement optimiste des estimations obtenues. Comme mentionné ci-dessus, lors de la production d'estimations au niveau local, les communautés vivant dans chaque zone disposent souvent d'informations (bien que subjectives) sur les valeurs plausibles des indicateurs en question, et les estimations fournies peuvent contredire ces connaissances locales. Ainsi, il est toujours nécessaire de rappeler à ceux qui utilisent des données statistiques que ces données comportent un certain degré d'erreur, et les mesures d'erreur accompagnant ces données devraient refléter les erreurs effectives relatives à chaque zone.

Des méthodes très développées pour l'estimation des erreurs quadratiques moyennes des estimateurs indirects correspondants ont également été décrites dans ce document. Cependant, on ne traite pas des mesures d'erreur incorporant des erreurs non dues à l'échantillonnage, telles que les

erreurs de couverture, les erreurs de non-réponse, les erreurs dans les données, l'imputation des données manquantes, etc. Ces questions doivent être étudiées plus avant dans le cadre de l'estimation sur petits domaines.

Ce document ne doit pas non plus être considéré comme un recueil exhaustif des méthodes de désagrégation des données (ou d'estimation des erreurs), car il existe un grand nombre de méthodes non décrites ici faute de place (voir Rao et Molina (2015) pour une description plus complète de la plupart des méthodes publiées précédemment). Ce document a seulement cherché à fournir une introduction au sujet considéré, en traitant les méthodes de base qui constituent les fondements de l'étude de méthodes plus avancées, mais avec une extension limitée à quelques-unes des méthodes les plus avancées qui sont conçues pour l'estimation d'indicateurs sur les conditions de vie.

Bibliographie

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015), Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67, 1-48.

Battese, G.E., Harter, R.M. and 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.

Bell, W. (1997), Models for county and state poverty estimates. Preprint, Statistical Research Division, U. S. Census Bureau.

Betti, G., Cheli, B., Lemmi, A. and Verma, V. (2006), Multidimensional and Longitudinal Poverty: an Integrated Fuzzy Approach, in Lemmi, A., Betti, G. (eds.) Fuzzy Set Approach to Multidimensional Poverty Measurement, 111-137, Springer, New York.

Casas-Cordero Valencia, C., Encina, J. and Lahiri, P. (2015), Poverty Mapping for the Chilean Comunas, In M. Pratesi (Ed.), Analysis of Poverty Data by Small Area Estimation: Methods for poverty mapping, New York: Wiley.

Correa, L., Molina, I., and Rao, J.N.K., (2012), Comparison of methods for estimation of poverty indicators in small areas. Unpublished report.

Datta, G.S., Fay, R.E. and Ghosh, M. (1991), Hierarchical and Empirical Bayes Multivariate Analysis in Small Area Estimation, in Proceedings of Bureau of the Census 1991 Annual Research Conference, U.S. Bureau of the Census, Washington, DC, 63-79.

Deville, J.C. and Särndal, C.E. (1992), Calibration estimation in Survey Sampling, Journal of the American Statistical Association, 87, 376-382.

Drew, D., Singh, M.P. and Choudhry, G.H. (1982), Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey, Survey Methodology, 8, 17-47.

Elbers, C., Lanjouw, J.O. and Lanjouw, P. (2003), Micro-level estimation of poverty and inequality. Econometrica, 71, 355-364.

Estevao, V., Hidiroglou, M. A. and Särndal, C. E. (1995), Methodological Principles for a Generalized Estimation Systems at Statistics Canada, Journal of Official Statistics, 11, 181-204.

Fay, R.E. (1987), Application of Multivariate Regression to Small Domain Estimation, in R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh (Eds.), Small Area Statistics, New York: Wiley, 91-102.

Fay, R.E. and Herriot, R.A. (1979), Estimation of Income for Small Places: An Application of James-Stein Procedures to Census Data, Journal of the American Statistical Association, 74, 269-277.
Ferretti, C. and Molina, I. (2012), Fast EB Method for Estimating Complex Poverty Indicators in Large Populations. Journal of the Indian Society of Agricultural Statistics, 66, 105-120.
Foster, J., Greer, J. and Thorbecke, E. (1984), A class of decomposable poverty measures, Econometrica, 52, 761-766.
Fuller, W.A. (1999), Environmental Surveys Over Time. Journal of Agricultural, Biological and Environmental Statistics, 4, 331-345. Regression Analysis for Sample Surveys, Sankhyā, Series C, 37, 117-132.
_____(1975), Regression Analysis for Sample Surveys, Sankhyā, Series C, 37, 117-132.
Graf, M., Marín, J.M. and Molina, I. (2018), A generalized mixed model for skewed distributions applied to small area estimation. Unpublished manuscript.
Ghosh, M. and Steorts, R.C. (2013), Two-stage benchmarking as applied to small area estimation, Test, 22, 670-687.
González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. and Santamaría, L. (2008), Bootstrap Mean Squared Error of a Small-Area EBLUP, Journal of Statistical Computation and Simulation, 75, 443-462.
_____(2007), Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model, Computational Statistics and Data Analysis, 51, 2720-2733.
González-Manteiga, W. (2010), Small area estimation under Fay-Herriot Models with nonparametric estimation of heteroscedasticity, Statistical Modelling, 10, 215-239.
Lumley, T. (2017), Survey: analysis of complex survey samples. R package version 3.32.
Molina, I. and Marhuenda, Y. (2015), sae: An R Package for Small Area Estimation, The R Journal, 7, 81-98.
Marhuenda, Y., Molina, I. and Morales, D. (2013), Small area estimation with spatio-temporal Fay-Herriot models. Computational Statistics and Data Analysis, 58, 308-325.
Marhuenda, Y., Molina, I., Morales, D., and Rao, J.N.K. (2018), Poverty mapping in small areas under a two-fold nested error regression model. Journal of the Royal Statistical Society, Series A, to be published shortly.
Molina, I. and Martín, N. (2018), Empirical best prediction under a nested error model with log transformation, Annals of Statistics, to be published shortly.
Molina, I. and Morales, D. (2009), Small area estimation of poverty indicators. Boletín de Estadística e Investigación Operativa (Bulletin of Statistics and Operations Research), 25, 218-225.
Molina, I., Nandram, B. and Rao, J.N.K. (2014), Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach. Annals of Applied Statistics, 8, 852-885.
Molina, I., and Rao, J.N.K. (2010), Small Area Estimation of Poverty Indicators. Canadian Journal of Statistics, 38, 369-385.
Molina, I., Salvati, N. and Pratesi, M. (2009), Bootstrap for estimating the MSE of the Spatial EBLUP. Computational Statistics, 24, 441-458.
Neri, L., Ballini, F. and Betti, G. (2005), Poverty and inequality in transition countries. Statistics in Transition, 7, 135-157.
Observatorio Social, Chilean Ministry of Social Development (2017), Metodología de estimación de pobreza a nivel comunal, con datos de Casen 2015. Aplicación de metodologías de estimación directa, de estimación para áreas pequeñas (SAE) e imputación de medias por conglomerados (IMC). (Poverty estimation methodology at the comuna level, with data from Casen 2015. Application of direct estimation methodologies, small area estimation (SAE) and cluster mean imputation (CMI)). Serie Documentos Metodológicos Casen (Casen Methodological Documents Series) 3428.
Pfeffermann, D. and Burk, L. (1990), Robust small area estimation combining time series and cross-sectional data. Survey Methodology, 16, 217-237.
Prasad, N.G.N. and 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.
Rao, J.N.K. and Molina (2015), Small area estimation, Second Ed., Hoboken, NJ: Wiley.
Rao, J.N.K. and Yu, M. (1992), Small area estimation by combining time series and cross-sectional data, Proceedings of the Section on Survey Research Method, American Statistical Association, 1-9.
Sen A. (1976), Poverty: An Ordinal Approach to Measurement. Econometrica, 44, 219-231.

Stukel, D. and Rao, J.N.K. (1999), On small-area estimation under two-fold nested error regression models. Journal of Statistical Planning and Inference, 78, 131-147.

Tillé, Y. and Matei, A. (2016), Sampling: Survey Sampling. R package version 2.8.

U.S. Bureau of Labor Statistics and U.S. Census Bureau. (2006), Design and Methodology: Current Population Survey, Technical Paper 66. Available at https://www.cen-sus.gov/prod/2006pubs/tp-66.pdf.

Torabi, M., and Rao, J.N.K. (2014), On small area estimation under a sub-area level model. Journal of Multivariate Analysis, 127, 36-55.

You, Y., and Rao, J.N.K. (2002a), A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights. Canadian Journal of Statistics, 30, 431-439.

_____(2002b), Small area estimation using unmatched sampling and linking models, Canadian Journal of Statistics, 30, 3-15.

Annexe

Résultats de l'estimation des revenus moyens et des taux de pauvreté à Montevideo

Tableau A1
Estimations directes, FH et EB-recensement du revenu moyen, erreurs quadratiques moyennes et coefficients de variation estimés de chaque estimateur, pour chaque secteur de recensement à Montevideo, pour les femmes
(En pesos uruguayens)

Sec-teur	n_d	Direct			FH			*Census EB* ***(EB-recensement)***		
		Est	var	cv	Est	eqm	cv	Est	eqm	cv
1	93	18 693,71	5057851,88	12,03	20 714,21	1240278,11	5,38	21 095,71	1042681,38	4,84
2	56	13 277,12	4664014,87	16,27	12 804,17	2589382,73	12,57	14 721,23	1423595,65	8,10
3	114	15 950,53	3705060,97	12,07	15 709,41	1449903,59	7,66	16 405,42	804002,00	5,47
4	172	21 964,73	3420289,14	8,42	22 818,88	1222004,22	4,84	23 513,94	685803,09	3,52
5	277	22 414,35	2388487,30	6,90	23 267,98	946571,10	4,18	22 041,95	423246,74	2,95
6	179	23 313,57	4128976,00	8,72	25 010,83	1352338,30	4,65	25 195,60	703110,23	3,33
7	421	23 755,31	1924432,84	5,84	22 592,03	1163899,22	4,78	20 071,35	256888,38	2,53
8	312	16 154,17	1263151,79	6,96	17 851,16	736152,51	4,81	18 028,97	321051,60	3,14
9	1 113	11 063,69	161639,01	3,63	11 127,05	151476,54	3,50	10 598,71	59822,24	2,31
10	3 482	22 823,33	230630,51	2,10	22 817,42	209158,99	2,00	20 764,66	24042,91	0,75
11	2 081	10 473,08	76660,70	2,64	10 390,16	74127,46	2,62	8 758,45	29006,70	1,94
12	1216	19 519,30	419289,74	3,32	19 251,93	347714,08	3,06	18 123,74	75945,76	1,52
13	1 844	10 741,54	95042,46	2,87	10 634,85	91443,15	2,84	8 652,12	38841,43	2,28
14	792	21 514,74	790097,23	4,13	21 340,52	560681,72	3,51	19 518,73	113098,81	1,72
15	607	25 157,71	1369375,39	4,65	24 472,89	931112,13	3,94	23 471,75	187914,34	1,85
16	960	7 748,40	84359,99	3,75	7 817,03	81443,96	3,65	8 592,14	68306,77	3,04
17	2 278	8 167,08	44950,84	2,60	8 217,67	44341,90	2,56	8 286,72	28268,28	2,03
18	2 227	34 942,88	746573,51	2,47	34 893,09	656698,10	2,32	33 015,11	64575,27	0,77
19	504	16 244,32	709726,52	5,19	15 953,47	526515,95	4,55	15 340,75	156341,38	2,58
20	1 402	12 724,70	168612,47	3,23	12 758,27	158968,95	3,13	12 436,10	65960,74	2,07
21	1 667	10 435,48	99478,80	3,02	10 526,60	95005,58	2,93	11 354,96	43098,30	1,83
22	1 073	14 104,97	272183,12	3,70	14 011,56	248767,70	3,56	13 370,10	82132,18	2,14
23	478	24 032,62	2084022,21	6,01	24 886,94	920424,29	3,85	23 547,61	230604,23	2,04
24	1 521	28 948,32	822681,41	3,13	28 302,65	588417,88	2,71	26 395,05	74187,48	1,03
99	1 364	9 614,82	101119,90	3,31	9 548,70	96674,58	3,26	9 081,88	47799,87	2,41

Source : D'après l'auteur.

Tableau A2
Estimations directes, FH et EB-recensement de la pauvreté non extrême (en pourcentage), erreurs quadratiques moyennes et coefficients de variation estimés de chaque estimateur, pour chaque secteur de recensement à Montevideo, pour les femmes

Secteur	n_d	Direct		FH		*Census EB* ***(EB-recensement)***	
		Est	var	Est	eqm	Est	eqm
1	93	0,00	0,0000	0,94	17,750	0,50	0,2707
2	56	0,00	0,0000	6,48	18,817	2,24	10,674
3	114	1,68	13,444	1,74	0,8148	1,51	0,4029
4	172	0,00	0,0000	0,00	15,588	0,30	0,2303
5	277	0,00	0,0000	0,00	15,516	0,42	0,1158
6	179	0,00	0,0000	0,00	17,350	0,20	0,1362
7	421	1,39	0,3732	1,12	0,3327	0,58	0,0810
8	312	1,06	0,3617	1,16	0,3075	0,98	0,1230
9	1 113	4,99	0,4874	5,80	0,3948	6,21	0,0888
10	3 482	2,41	0,0813	2,43	0,0786	0,55	0,0184
11	2 081	11,40	0,6827	10,79	0,5181	10,55	0,0393
12	1 216	0,88	0,0681	0,92	0,0662	0,90	0,0341
13	1 844	12,85	0,9809	12,03	0,6817	10,97	0,0521
14	792	1,80	0,2872	1,82	0,2580	0,70	0,0532
15	607	0,20	0,0406	0,20	0,0403	0,29	0,0480
16	960	9,75	0,9567	10,43	0,6890	11,09	0,1267
17	2 278	11,97	0,5287	11,84	0,4425	12,28	0,0449
18	2 227	0,27	0,0128	0,26	0,0127	0,06	0,0148
19	504	0,69	0,1549	0,83	0,1450	1,72	0,0783
20	1 402	6,64	0,5047	6,27	0,4047	3,83	0,0518
21	1 667	5,61	0,3258	5,89	0,2824	5,08	0,0487
22	1 073	4,30	0,4345	4,11	0,3600	2,89	0,0497
23	478	0,00	0,0000	0,00	15,526	0,29	0,0510
24	1 521	0,17	0,0136	0,17	0,0135	0,17	0,0233
99	1 364	10,98	0,8465	10,43	0,6014	9,65	0,0898

Source : D'après l'auteur.

Tableau A3
Estimations directes, FH et EB-recensement du revenu moyen, erreurs quadratiques moyennes et coefficients de variation estimés de chaque estimateur, pour chaque secteur de recensement à Montevideo, pour les hommes
(En pesos uruguayens)

Secteur	n_d	Direct			FH			*Census EB* ***(EB-recensement)***		
		Est	var	cv	Est	eqm	cv	Est	eqm	cv
1	74	24 836,71	11478551,50	13,64	23 153,44	2192630,31	6,40	25 762,01	1516931,82	4,78
2	65	20 460,87	16472932,99	19,84	15 443,37	2435603,57	10,11	15 765,25	1269312,32	7,15
3	72	14 298,89	4973475,40	15,60	17 664,10	1405480,47	6,71	16 069,72	1460721,97	7,52
4	147	26 634,78	6878201,92	9,85	25 124,25	1749062,81	5,26	23 452,91	883029,38	4,01
5	218	23 222,64	3977409,63	8,59	24 597,89	1109370,86	4,28	21 535,17	568700,05	3,50
6	141	28 783,87	8278641,51	10,00	29 394,58	2072316,24	4,90	26 974,58	956465,19	3,63
7	343	25 127,32	2855763,61	6,73	25 322,16	1257216,13	4,43	20 211,77	353216,26	2,94
8	261	17 701,92	2187547,36	8,36	19 904,65	1018640,94	5,07	17 975,03	384896,60	3,45
9	1 032	11 475,41	190257,67	3,80	11 517,36	174746,99	3,63	10 253,44	84239,24	2,83
10	2 820	24 575,73	396689,08	2,56	24 658,6	348365,27	2,39	21 279,59	39334,03	0,93
11	1 882	11 079,86	99418,16	2,85	10 975,09	95447,04	2,81	9 008,44	42747,93	2,30
12	1 009	21 868,24	638076,52	3,65	21 342,01	481650,83	3,25	18 731,09	88898,30	1,59
13	1 712	10 897,03	106703,48	3,00	10 766,13	103066,30	2,98	8 566,47	44008,29	2,45
14	641	22 605,70	1090803,69	4,62	22 636,34	664159,72	3,60	19 332,12	177272,29	2,18
15	509	28 797,75	2175732,47	5,12	27 945,27	1339582,93	4,14	23 540,89	232540,09	2,05
16	894	8 920,20	130401,00	4,05	8 893,51	125439,87	3,98	9 342,98	73446,28	2,90
17	2 095	8 749,60	58161,78	2,76	8 830,39	57265,48	2,71	8 402,07	37470,56	2,30
18	1 723	38 931,89	1332962,25	2,97	38 347,54	1019441,37	2,63	33 874,56	93681,13	0,90
19	417	17 855,77	1230524,64	6,21	17 640,61	755965,09	4,93	15 893,35	237964,08	3,07
20	1 179	13 531,48	248104,13	3,68	13 611,71	229000,88	3,52	12 318,07	72710,30	2,19
21	1 498	11 147,80	132574,18	3,27	11 290,17	125020,79	3,13	11 380,39	65707,49	2,25
22	929	15 394,11	397876,11	4,10	15 137,23	349742,04	3,91	13 156,00	92731,14	2,31
23	392	27 941,71	3640071,50	6,83	28 804,9	1294705,53	3,95	23 483,98	321582,84	2,41
24	1 170	29 940,63	1097458,14	3,50	29 943,72	745397,44	2,88	26 410,63	108953,03	1,25
99	1 241	10 230,59	130610,91	3,53	10 227,46	124563,28	3,45	9 379,01	67878,14	2,78

Source : D'après l'auteur.

Tableau A4
Estimations directes, FH et EB-recensement de la pauvreté non extrême (en pourcentage), erreurs quadratiques moyennes et coefficients de variation estimés de chaque estimateur, pour chaque secteur de recensement à Montevideo, pour les hommes.

Secteur	n_d	Direct		FH		*Census EB* ***(EB-recensement)***	
		Est	var	Est	eqm	Est	eqm
1	74	0,00	0,0000	0,00	23,526	0,24	0,4195
2	65	1,89	34,599	4,06	14,112	1,94	11,930
3	72	1,10	11,480	0,82	0,8380	1,90	0,7071
4	147	0,80	0,6193	0,30	0,5133	0,37	0,1798
5	218	0,68	0,4522	0,66	0,3936	0,57	0,1535
6	141	0,00	0,0000	0,00	23,300	0,19	0,2299
7	343	1,39	0,4707	1,27	0,4145	0,73	0,1259
8	261	0,00	0,0000	1,89	20,897	1,15	0,1482
9	1 032	5,32	0,5585	6,04	0,4740	7,40	0,1099
10	2 820	2,34	0,0959	2,38	0,0928	0,61	0,0233
11	1 882	10,23	0,7013	9,86	0,5710	10,38	0,0589
12	1 009	0,31	0,0312	0,34	0,0309	0,93	0,0605
13	1 712	12,81	10,956	11,88	0,8397	11,89	0,0636
14	641	1,99	0,3835	2,09	0,3381	0,88	0,0724
15	509	0,53	0,1388	0,50	0,1337	0,36	0,0644
16	894	9,07	0,9783	9,60	0,7497	9,33	0,1292
17	2 095	12,09	0,5981	11,83	0,5141	12,47	0,0419
18	1 723	0,17	0,0134	0,16	0,0134	0,08	0,0199
19	417	0,83	0,2247	0,98	0,2099	1,72	0,1449
20	1 179	7,04	0,6212	6,60	0,5067	4,28	0,0596
21	1 498	6,36	0,4374	6,47	0,3789	5,45	0,0606
22	929	6,35	0,8152	5,60	0,6296	3,39	0,0734
23	392	0,71	0,2482	0,73	0,2305	0,39	0,0902
24	1 170	0,21	0,0209	0,22	0,0208	0,24	0,0355
99	1 241	10,76	0,9442	10,44	0,7215	9,38	0,0935

Source : D'après l'auteur.

NATIONS UNIES

CEPALC

Séries

Études statistiques

Numéros publiés

Une liste complète ainsi que des fichiers pdf sont disponibles à l'adresse suivante www.eclac.org/publicaciones

97. Désagrégation des données pour les enquêtes auprès des ménages: utilisation de méthodes d'estimation sur petits domaines, Isabel Molina (LC/TS.2018/82/Rev.1), 2022.
96. ¿Cuál es el alcance de las transferencias no contributivas en América Latina?: discrepancias entre encuestas y registros, Pablo Villatoro, Simone Cecchini (LC/TS.2018/46), 2018.
95. Avances y desafíos de las cuentas económico-ambientales en América Latina y el Caribe, Franco Carvajal (LC/TS.2017/148), 2018.
94. La situación de las estadísticas, indicadores y cuentas ambientales en América Latina y el Caribe (LC/TS.2017/135), 2017.
93. Indicadores no monetarios de carencias en las encuestas de los países de América Latina: disponibilidad, comparabilidad y pertinencia, Pablo Villatoro (LC/TS.2017/130), 2017.
92. Un índice de pobreza multidimensional para América Latina, María Emma Santos, Pablo Villatoro, Xavier Mancero Pascual Gerstenfeld (LC/L.4129), 2015.
91. Ajuste de los ingresos de las encuestas a las Cuentas Nacionales. Una revisión de la literatura, Pablo Villatoro (LC/L.4002), 2015.
90. La evolución del ingreso de los hogares en América Latina durante el período 1990-2008 ¿Ha sido favorable a los pobres?, Fernando Medina y Marco Galván (LC/L.3975), 2015.
89. ¿Qué es el crecimiento propobre?, Fundamentos teóricos y metodologías para su medición, Fernando Medina y Marco Galván (LC/L.3883), 2014.
88. Cuentas satélite y cuentas de salud: un análisis comparativo, Federico Dorin, Salvador Marconi y Rafael Urriola (LC/L.3865), 2014.

ÉTUDES STATISTIQUES

Numéros publiés:

97 Désagrégation des données pour les enquêtes auprès des ménages
Utilisation de méthodes d'estimation sur petits domaines
Isabel Molina

96 ¿Cuál es el alcance de las transferencias no contributivas en América Latina?
Discrepancias entre encuestas y registros
Pablo Villatoro y Simone Cecchini

95 Avances y desafíos de las cuentas económico-ambientales en América Latina y el Caribe
Franco Carvajal

94 La situación de las estadísticas, indicadores y cuentas ambientales en América Latina y el Caribe

CEPALC
Commission économique pour l'Amérique latine et les Caraïbes (CEPALC)
Comisión Económica para América Latina y el Caribe (CEPAL)
www.cepal.org

LC/TS.2018/82/REV.1