Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Estimation sur petits domaines portant sur des chiffres pondérés d'enquête dans un modèle spatial à niveau agrégé

par Hukum Chandra, Ray Chambers et Nicola Salvati

Date de diffusion : le 7 mai 2019

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Publication autorisée par le ministre responsable de Statistique Canada

© Sa Majesté la Reine du chef du Canada, représentée par le ministre de l'Industrie 2019

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada.

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, 2019 (numéro spécial)
Vol. 45, nº 1, pp. 33-63
Statistique Canada, nº 12-001-X au catalogue

Estimation sur petits domaines portant sur des chiffres pondérés d'enquête dans un modèle spatial à niveau agrégé

Hukum Chandra, Ray Chambers et Nicola Salvati[1]

Résumé

Le prédicteur empirique dans une version au niveau du domaine du modèle linéaire généralisé mixte (MLGM) est amplement employé dans une estimation sur petits domaines (EPD) portant sur des dénombrements, mais cette méthode ne fait appel ni à la pondération d'échantillonnage ni à l'information de mise en grappes, lesquelles sont essentielles à une inférence valable avec les échantillons informatifs produits par les plans de sondage complexes d'aujourd'hui. Nous décrirons une méthode d'EPD qui intègre cette information d'échantillonnage à l'estimation de proportions ou de chiffres de petits domaines dans une version au niveau du domaine du MLGM. Nous élargissons encore notre méthode en employant une version avec dépendance spatiale du MLGM (MLGMS). Il est aussi question de l'estimation de l'erreur quadratique moyenne (EQM) pour cette méthode. Nous appliquons ensuite le traitement d'EPD à l'estimation de la fréquence de la pauvreté des ménages dans divers districts ruraux de l'État d'Uttar Pradesh en Inde, et ce, en couplant, d'une part, les données de l'enquête 2011-2012 sur les dépenses de consommation des ménages recueillies par le *National Sample Survey Office* (NSSO) et, d'autre part, les données du recensement de 2011 en Inde. Les résultats de cette application font voir un gain appréciable de précision avec les nouvelles méthodes comparativement aux estimations directes d'enquête.

Mots-clés : Enquêtes complexes; estimateur direct pondéré d'enquête; estimation de la pauvreté; modèle spatial; cartographie.

1 Introduction

On planifie généralement les enquêtes par sondage de manière à produire des estimations des caractéristiques d'intérêt de la population surtout à des niveaux géographiques supérieurs (pays et ses états constitutifs, par exemple). La taille de l'échantillon est fixée pour que les estimateurs directs d'un grand domaine livrent des estimations fiables. Par estimateurs directs, nous entendons des estimateurs qui exploitent uniquement des données d'échantillon pondérées du domaine d'intérêt. Dans maintes situations pratiques cependant, le but est d'estimer les paramètres de domaines comptant seulement un petit nombre d'observations d'échantillon. Nous employons le terme « petits domaines » pour désigner des domaines où la taille d'échantillon ne suffit pas à des estimations directes de la précision voulue. Là où l'estimation directe est impossible, on doit se rabattre sur d'autres méthodes par modèle pour des estimations sur petits domaines. Cet emploi dépend de la spécification du modèle utilisé et de la disponibilité au niveau de la population de renseignements auxiliaires relatifs à la variable d'intérêt, ce qu'on qualifie communément de méthodes indirectes (Rao, 2003). La théorie qui le sous-tend est celle de l'estimation sur petits domaines (EPD) dont les techniques visent à produire des estimations fiables avec de petites tailles d'échantillon grâce à un modèle « couplant » les petits domaines pour « emprunter de la puissance » aux données d'échantillon d'autres petits domaines (voir, par exemple, Pfeffermann (2002) et Rao et Molina (2015)). Dans ce contexte, nous distinguons les méthodes d'EPD par modèle au niveau des unités et au niveau des domaines. Dans le

1. Hukum Chandra, Indian Agricultural Statistics Research Institute, avenue Library, New Delhi, Inde. Courriel : hchandra12@gmail.com; Ray Chambers, Université de Wollongong, Australie; Nicola Salvati, Université de Pise, Italie.

premier cas, le modèle vise des mesures individuelles d'enquête avec des effets de domaine et, dans le second, il sert à lisser la variabilité en situation d'instabilité des estimations directes d'enquête au niveau du domaine. La modélisation au niveau du domaine a normalement sa place lorsqu'on ne dispose pas de données au niveau des unités ou, comme c'est souvent le cas, là où on ne peut obtenir de covariables de modèle (variables de recensement, par exemple) qu'au niveau du domaine. Le modèle de Fay-Herriot (Fay et Herriot, 1979) est un modèle répandu au niveau du domaine où on suppose que des estimations d'enquête par domaine sont disponibles et qu'elles suivent un modèle linéaire mixte au niveau du domaine avec des effets aléatoires de domaine indépendants. Un tel modèle peut aussi recevoir une pondération d'enquête dans un traitement d'EPD par des estimations directes pondérées d'enquête dans l'ajustement du modèle linéaire mixte. Si la variable d'intérêt est discontinue (consistant, par exemple, en données binaires ou données de dénombrement), un modèle linéaire généralisé mixte (MLGM) est souvent employé. Si la variable d'intérêt est binaire et que l'inférence porte sur une proportion de petits domaines, on fait couramment appel à la fonction de lien logistique (aussi appelée modèle linéaire mixte logistique). Un prédicteur plug-in empirique (PE) sert habituellement à l'estimation de proportions de petits domaines avec la méthode MLGM (voir, par exemple, Chandra, Chambers et Salvati (2012), Salvati, Chandra et Chambers (2012), Rao et Molina (2015) et les références présentées), bien que ce ne soit pas là le prédicteur le plus efficace avec un tel modèle. Une solution de rechange est le meilleur prédicteur empirique (Jiang, 2003) dont la forme n'est pas fermée en général et qui doit se calculer par approximation numérique, ce qui n'est pas toujours simple. C'est pourquoi les organismes statistiques nationaux tendent à privilégier le calcul d'une approximation comme le PE. Nous savons par ailleurs que des prédicteurs du type PE sont utilisés dans Molina, Saei et José Lombardìa (2007) et Lopez-Vizcaino, Lombardìa et Morales (2013). S'il n'y a de disponibles que des données au niveau du domaine, on peut envisager une version domaine du MLGM pour le traitement d'EPD (voir, par exemple, Saei et Chambers (2003), Johnson, Chandra, Brown et Padmadas (2010), Chandra, Salvati et Sud (2011), Chandra, Salvati et Chambers (2017) et les références présentées). À la différence du modèle de Fay-Herriot, ce modèle pose implicitement un échantillonnage aléatoire simple avec remise dans chaque domaine et écarte la pondération d'enquête. Cela risque malheureusement de biaiser sérieusement les estimations en cas de grave déséquilibre des échantillons de petits domaines en ce qui concerne les caractéristiques clés de la population. Il semblerait donc que la pondération d'enquête est inévitable si on désire produire des estimations sur petits domaines représentatives.

La modélisation d'estimations pondérées d'enquête de proportions dans une distribution continue comme dans le modèle de Fay-Herriot peut poser un problème, puisque la distribution d'échantillonnage des proportions en question est foncièrement discrète à cause des petites tailles d'échantillon et du caractère binaire des observations sous-jacentes. Il se peut, par exemple, qu'un traitement par modèle linéaire mène à des prédictions négatives ou, plus probablement encore, à des intervalles de prédiction comprenant des valeurs négatives, autant de résultats absurdes. Il est clair que les MLGM se prêtent davantage à la modélisation de données foncièrement discrètes venant d'estimations directes pondérées d'enquête (voir Liu, Lahiri et Kalton (2014), Ghosh, Natarajan, Stroud et Carlin (1998) et Rao (2003, sections 5.6 et 10.11). Malec, Sedransk, Moriarity et LeClere (1997) décrivent un modèle hiérarchique de Bayes appliqué à des

données binaires d'enquête. Ces auteurs ont examiné la pondération d'enquête comme covariable dans le modèle pour ne constater aucune amélioration dans leur exemple de données au niveau des comtés de la *National Health Interview Survey*. Si on la compare à l'estimation de paramètres linéaires pondérés selon l'enquête comme les moyennes et les totaux de petits domaines, l'estimation de proportions ou de chiffres pondérés selon l'enquête dans des modèles de petits domaines au niveau du domaine n'a guère suscité de travaux de recherche. Récemment, des auteurs comme Mercer, Wakefield, Chen et Lumley (2014), Liu et coll. (2014) et Franco et Bell (2013) ont décrit comment il était possible de recourir à une pondération d'enquête dans un modèle hiérarchique bayésien d'estimation de proportions de petits domaines en employant des tailles d'échantillon efficaces pour l'inférence d'une distribution binomiale sous-jacente de manière à préserver les variances d'échantillonnage estimées par une fonction de variance généralisée. Ces traitements se définissent dans un cadre bayésien. Dans le présent article, nous adoptons une méthode fréquentiste et exposons une méthode d'EPD utilisant les poids d'enquête pour les données binaires lors de l'estimation de proportions de petits domaines dans une version MLGM au niveau du domaine. D'après Korn et Graubard (1998), nous modélisons les estimations pondérées d'enquête comme proportions binomiales en choisissant une « taille d'échantillon efficace » permettant de faire correspondre la variance binomiale à la variance d'échantillonnage des estimations. Utiliser une taille d'échantillon efficace plutôt que la taille d'échantillon réelle est admettre l'information variable de chaque domaine dans un échantillonnage complexe. De plus, nous supposons normalement que les effets aléatoires par domaine sont indépendants dans le traitement MLGM, c'est-à-dire que les petits domaines qui entre en jeu sont indépendants les uns des autres. Il reste que, dans la pratique, les limites de petits domaines sont le plus souvent arbitraires et qu'on ne voit pas vraiment pourquoi les unités d'un côté d'une limite ne devraient pas être généralement en corrélation avec les unités de l'autre côté (Pratesi et Salvati, 2008). Un moyen d'intégrer une telle information spatiale à la modélisation d'EPD est d'élargir le modèle des effets aléatoires pour admettre des effets de domaine en corrélation spatiale dans un modèle autorégressif simultané (ARS), par exemple (voir Anselin (1992) et Cressie (1993)). Ce sont des modèles qui introduisent un comportement spatial dans les effets aléatoires du modèle, qui sont habituellement les effets de domaine dans le contexte d'EPD. Singh, Shukla et Kundu (2005), Pratesi et Salvati (2008), Pratesi et Salvati (2009), Molina, Salvati et Pratesi (2009), Marhuenda, Molina et Morales (2013) et Porter, Wikle et Holan (2015) ont envisagé des applications des modèles ARS à l'estimation sur petits domaines. Ces modèles servent communément à la modélisation de la dépendance spatiale dans un traitement d'EPD fréquentiste. De leur côté, les traitements d'EPD bayésiens privilégient les modèles autorégressifs conditionnels (ARC) pour cette modélisation (voir Besag, York et Mollié (1991), Leroux, Lei et Breslow (1999) et Mercer et coll. (2014)). Nous optons pour le traitement d'EPD fréquentiste et proposons un élargissement du MLGM au niveau du domaine pour tenir compte de la dépendance spatiale entre petits domaines par une spécification ARS. Nous établissons dans ce modèle des estimations sur petits domaines qui font intervenir l'information de plan de sondage.

Voici comment nous structurons notre propos. La section 2 décrit les données de l'enquête 2011-2012 sur les dépenses de consommation des ménages du *National Sample Survey Office* (NSSO) et du recensement 2011 de la population en Inde pour l'estimation de la fréquence de la pauvreté des ménages

dans les districts ruraux de l'État indien d'Uttar Pradesh. À la section 3, nous exposons le contexte théorique de la version au niveau du domaine du MLGM, puis examinons l'emploi d'une taille d'échantillon efficace permettant d'intégrer l'information de la pondération d'enquête au moment d'estimer les proportions de petits domaines avec ce modèle. Nous présentons également dans cette section l'extension à la corrélation spatiale du MLGM au niveau du domaine, ainsi que le prédicteur empirique fondé sur ce modèle. À la section 4, nous présentons les résultats de l'application à la fréquence de la pauvreté avec diverses mesures de diagnostic et une carte de la pauvreté démontrant les inégalités entre districts de la répartition des ménages ruraux pauvres de l'État d'Uttar Pradesh. À la section 5 enfin, nous résumons notre propos et livrons des observations en conclusion.

2 Description des données

Dans cette section, nous présentons les sources principales de données, à savoir l'enquête sur les dépenses de consommation des ménages (EDCM) du NSSO de 2011-2012 dans les régions rurales de l'État d'Uttar Pradesh et le recensement 2011 de la population en Inde dans l'application aux petits domaines que nous allons examiner. Les données de ces sources serviront à estimer la proportion de ménages pauvres au niveau des districts de l'Uttar Pradesh. Cet État est le plus populeux du pays et compte environ 16,16 % de la population indienne. Il s'étend sur 243 290 kilomètres carrés, soit 6,88 % de la superficie totale du pays. En Inde, on produit des estimations de la pauvreté pour tous les États en séparant les secteurs ruraux des secteurs urbains. Dans notre analyse, nous nous limiterons à parler des régions rurales de l'Uttar Pradesh, car une proportion approximative de 78 % des habitants de cet État se trouvent en milieu rural d'après le recensement de la population de 2011. Le NSSO mène régulièrement des enquêtes nationales sur les dépenses de consommation des ménages dans des cycles d'une durée normale d'un an. Ces enquêtes visent à produire des estimations des dépenses de consommation des ménages en moyenne mensuelle par personne (DCMP), de la distribution des ménages et des personnes à l'échelle des DCMP et de la ventilation des DCMP moyennes par groupe de produits et selon les secteurs ruraux et urbains du pays pour les États et les territoires de l'Union et différents groupes socioéconomiques. Ces indicateurs comptent parmi les mesures les plus importantes des conditions de vie dans les domaines de population considérés. Les enquêtes en question se font par interviews auprès d'un échantillon représentatif de ménages choisis au hasard par un plan de sondage approprié pour toute l'étendue du pays. Précisons que le plan de sondage de cette enquête du NSSO crée un échantillonnage aléatoire stratifié à plusieurs degrés avec les districts comme strates, les villages comme unités du premier degré et les ménages comme unités du second degré. Bien que de tels sondages livrent des estimations fiables et représentatives au niveau des États et du pays, ils ne peuvent directement servir à produire des estimations sûres au niveau des districts à cause des petites tailles d'échantillon. Le district est un domaine fort important du processus de planification en Inde, mais aucune enquête ne se propose de produire des estimations à ce niveau. L'absence de mesures de résultats solides et fiables au niveau des districts impose des contraintes à la conception des interventions ciblées et à

l'élaboration des politiques. Ce qui est plus important encore, c'est que les estimations étatiques et nationales ne prennent pas bien la mesure de l'ampleur des inégalités géographiques, d'où une moindre portée de l'évaluation des progrès locaux dans et entre les districts. Il faut bien avouer par ailleurs qu'une enquête au niveau des districts coûterait très cher et prendrait beaucoup de temps. Dans l'EDCM, on a interrogé 5 916 ménages au total dans 71 districts de l'État d'Uttar Pradesh. La taille d'échantillon dans les districts variait de 32 à 128 pour une moyenne de 83. Il est évident que les tailles d'échantillon sont relativement modestes au niveau des districts, le taux moyen d'échantillonnage s'établissant à 0,0002. Il est donc difficile dans le cadre de cette enquête de produire des estimations directes d'enquête pour les districts avec les erreurs-types liées. Il est possible de résoudre ce problème de petites tailles d'échantillon par la méthodologie d'EPD à condition de disposer d'une information auxiliaire pour étoffer les données d'échantillon limitées émanant d'un district (Rao et Molina, 2015; Tzavidis, Salvati, Pratesi et Chambers, 2008).

On notera ici que la variable visée Y est binaire au niveau des unités (ménages) dans le fichier publié des données d'enquête. La variable indique si un ménage est pauvre ou non. Dans notre application cependant, nous nous employons à produire des estimations lorsque les données disponibles sont des données de dénombrement des ménages pauvres dans l'échantillon de chaque district. Dans ce contexte, un ménage dont les DCMP se situent sous le seuil de la pauvreté de l'État est défini comme pauvre. Le seuil de pauvreté employé dans cette étude (768 roupies) est celui qu'a fixé la Planning Commission du gouvernement indien pour 2011-2012. Le paramètre d'intérêt est alors la proportion de ménages ruraux pauvres dans chaque district. Les variables auxiliaires (covariables) de notre analyse sont tirées du recensement indien de la population de 2011. Elles sont seulement disponibles en dénombrements au niveau des districts, d'où la nécessité, comme nous le décrirons dans la prochaine section, d'appliquer des méthodes d'EPD par modèle de petits domaines au niveau du domaine pour établir les estimations correspondantes. Une cinquantaine de covariables sont disponibles pour l'analyse d'EPD. Nous avons donc procédé à une analyse préliminaire des données pour définir des covariables appropriées de modélisation d'EPD. Nous avons effectué une analyse en composantes principales (ACP) afin de dégager des valeurs composites pour des groupes choisis de variables. Plus précisément, nous avons soumis séparément à l'ACP trois groupes de variables, tous mesurés au niveau des districts et désignés par S1, S2 et S3 dans la suite du texte. Le premier groupe (S1) consistait en taux d'alphabétisation et en proportions de la population de travailleurs selon le sexe. La première composante principale (S11) pour ce groupe expliquait 51 % de la variabilité de S1 et, si on ajoutait l'autre composante (S12), on obtenait 85 %. Le deuxième groupe (S2) consistait en proportions selon le sexe de travailleurs, de cultivateurs et d'ouvriers agricoles à titre principal. La première composante principale (S21) pour ce groupe expliquait 49 % de la variabilité de S2 et, avec l'autre composante (S22), 67 %. Le troisième groupe (S3) consistait en proportions selon le sexe de cultivateurs et d'ouvriers agricoles marginaux. La première composante principale (S31) pour ce groupe expliquait 61 % de la variabilité de S3 et, si on ajoutait l'autre composante (S32), on obtenait 78 %. À l'aide des méthodes décrites plus en détail dans les sections qui suivent, nous avons ajusté un modèle linéaire généralisé avec les estimations directes d'enquête des proportions de ménages ruraux pauvres comme variable de réponse et les six résultats de

l'analyse en composantes principales (S11, S12, S21, S22, S31 et S32) comme covariables possibles. Le modèle finalement retenu comprenait les trois covariables S11, S21 et S31 avec une déviance résiduelle et des valeurs AIC respectives de 327,18 et 636,89. Ce modèle a ensuite servi à des estimations par district de la fréquence de la pauvreté en milieu rural, c'est-à-dire à des estimations de proportion de pauvres au niveau des districts.

Un grand problème lié à l'application de l'EPD dans un grand nombre de pays en développement ou sous-développés est que les données administratives ou les données d'état civil pouvant donner des covariables dans le modèle d'EPD ne sont pas disponibles au niveau des petits domaines. Et si des variables auxiliaires sont disponibles (totalisations du recensement, par exemple), elles ne sont normalement pas suffisamment en association avec la variable d'intérêt. L'information reste donc fort limitée pour la production d'estimations sur petits domaines, auquel cas il peut être bon de compléter les données disponibles en utilisant une information géospatiale sur les petits domaines. Dans notre étude, nous nous reportons à la situation géographique (par les centroïdes) des divers districts de l'État indien d'Uttar Pradesh pour élargir notre modèle d'EPD et ainsi mettre en corrélation les dénombrements de ruraux pauvres entre districts voisins.

3 Estimation sur petits domaines avec le MLGM au niveau du domaine

3.1 Effets aléatoires de domaine sans corrélation spatiale

Nous supposons qu'un échantillonnage probabiliste permet de tirer un échantillon s de taille n d'une population finie U de taille N consistant en D domaines disjoints U_i $(i = 1, \ldots, D)$. Comme c'est la règle, nous parlerons ici de petits domaines ou seulement de domaines. Nous posons également qu'il y a un nombre connu N_i d'unités de population dans le petit domaine i et que n_i de ces unités sont échantillonnées. Le nombre total d'unités de la population est $N = \sum_{i=1}^{D} N_i$, et la taille totale d'échantillon correspondante, $n = \sum_{i=1}^{D} n_i$. Soit s la collection d'unités de l'échantillon et s_i le sous-ensemble tiré du petit domaine i $(|s_i| = n_i)$, nous employons des expressions comme $j \in i$ et $j \in s$ pour les unités composant le petit domaine i et l'échantillon s respectivement. De même, r_i désigne l'ensemble d'unités du petit domaine i qui ne sont pas dans l'échantillon avec $|r_i| = N_i - n_i$ et $U_i = s_i \cup r_i$. Soit y_{ij} la valeur de la variable d'intérêt pour l'unité j $(j = 1, \ldots, N_i)$ dans le domaine i. La variable d'intérêt aux valeurs y_{ii} est binaire ($y_{ij} = 1$ si le ménage j dans le domaine i est pauvre et 0 dans les autres cas). Le but est d'estimer les chiffres de population de petits domaines $y_i = \sum_{j \in U_i} y_{ij}$, ou, comme équivalent, la proportion de petits domaines $P_i = N_i^{-1} \sum_{j \in U_i} y_{ij}$, dans le domaine i $(i = 1, \ldots, D)$. L'estimateur direct type d'enquête (DIR dans la suite du texte) pour P_i est $p_{iw} = \sum_{j \in s_i} \tilde{w}_{ij} y_{ij}$, où $\tilde{w}_{ij} = w_{ij} \Big/ \sum_{j \in s_i} w_{ij}$ est le poids d'échantillon normalisé pour l'unité j dans le domaine i avec $\sum_{j \in s_i} \tilde{w}_{ij} = 1$ et où w_{ij} est le poids

d'enquête pour l'unité j dans le domaine i. On approche la variance de plan de sondage estimée de DIR par $v(p_{iw}) \approx \sum_{j \in s_i} \tilde{w}_{ij} (\tilde{w}_{ij} - 1)(y_{ij} - p_{iw})^2$. On emploie cette formule pour l'estimateur de variance de DIR d'après Särndal, Swensson et Wretman (1992; voir pages 43, 185 et 391), et après les simplifications $w_{ij} = a_{ij}^{-1}$, $a_{ij,ij} = a_{ij}$ et $a_{ij,ik} = a_{ij}a_{ik}$, $j \neq k$, où a_{ij} est la probabilité d'inclusion de premier ordre de l'unité j dans le domaine i et où $a_{ij,ik}$ est la probabilité d'inclusion de second ordre des unités j et k dans le domaine i. Dans un échantillonnage aléatoire simple (EAS), $w_{ij} = N_i n_i^{-1}$ et DIR est alors $p_i = n_i^{-1} y_{si}$, avec une variance estimée $v(p_i) \approx n_i^{-1} p_i (1 - p_i)$, où $y_{si} = \sum_{j \in s_i} y_{ij}$ désigne le chiffre de dénombrement d'échantillon dans le domaine i.

Si le plan de sondage est informatif, la version EAS de DIR définie au paragraphe précédent peut être entachée d'un biais. Dans le cas de l'EDCM 2011-2012, nous avons donc calculé des effets de plan de sondage par district définis comme le rapport entre la variance des estimations pondérées (pour un plan donné d'échantillonnage) et la variance des estimations non pondérées (pour un échantillonnage aléatoire simple). On pouvait constater que ces effets étaient supérieurs à l'unité dans tous les districts sauf trois, les valeurs relevées variant de 1,17 à 8,44 pour une moyenne de 2,71. On est nettement porté à penser que le plan d'échantillonnage de l'EDCM de 2011-2012 est informatif. Ajoutons que l'estimateur direct d'enquête DIR repose sur des données d'échantillon par domaine et peut donc se révéler fort imprécis si la taille d'échantillon par domaine est petite, voire impossible à calculer si cette même taille est nulle. Pour résoudre ce problème, nous pouvons néanmoins employer des méthodes d'EPD par modèle qui « empruntent de la puissance » grâce à un modèle statistique commun pour tous les petits domaines en question (voir Rao et Molina (2015)).

Posons maintenant que les données disponibles consistent en agrégats d'échantillon y_{si} (dénombrements d'échantillon des ménages pauvres) avec les valeurs de covariables de contexte par domaine. Ainsi, nous obtenons le chiffre de dénombrement y_{si} pour le domaine i avec un vecteur k de covariables $\mathbf{x}_i$ par domaine venant de sources d'information secondaires (recensement ou registres administratifs, par exemple). Si nous ne tenons pas compte du plan d'échantillonnage, nous pouvons supposer que le dénombrement d'échantillon y_{si} suit une distribution binominale avec les paramètres n_i et π_i, soit $y_{si} \sim \text{Bin}(n_i, \pi_i)$, où π_i est la probabilité d'occurrence pour une unité de population dans le domaine i ou la probabilité de fréquence dans ce même domaine. D'après Saei et Chambers (2003), Johnson et coll. (2010) et Chandra et coll. (2011), ce modèle couplant la probabilité π_i et les covariables $\mathbf{x}_i$ est le modèle linéaire mixte logistique de la forme

$$\text{logit}(\pi_i) = \ln\left\{\pi_i (1 - \pi_i)^{-1}\right\} = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i, \tag{3.1}$$

avec $\pi_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta} + u_i)\{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta} + u_i)\}^{-1} = \text{expit}(\mathbf{x}_i^T \boldsymbol{\beta} + u_i)$. Là, $\boldsymbol{\beta}$ est le vecteur k de coefficients de régression, ce qu'on appelle souvent le vecteur des effets fixes, et u_i est l'effet aléatoire par

domaine avec $u_i \sim N(0, \sigma_u^2)$. Le modèle (3.1) est un cas d'espèce du modèle linéaire généralisé mixte (MLGM) avec une fonction de lien logits (Breslow et Clayton, 1993). Nous pouvons observer que les paramètres $\boldsymbol{\beta}$ et σ_u^2 sont les mêmes pour chaque domaine, c'est-à-dire qu'ils peuvent s'estimer à l'aide des données de tous les petits domaines. C'est ce que l'on fait habituellement en « empilant » les modèles D au niveau du domaine en (3.1) pour produire au niveau de la population un modèle de la forme

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \tag{3.2}$$

avec $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_D)^T = \text{expit}(\boldsymbol{\eta})$, où $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_D)^T$, $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_D^T)^T$ est la matrice $D \times k$ des covariables, où $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_D)^T$ est une matrice de covariables connues de dimension $D \times D$ caractérisant les différences entre les petits domaines, où $\mathbf{z}_i$ est le vecteur $D\,(0, \ldots, 1, \ldots, 0)^T$ avec le 1 dans la i^{e} position et où $\mathbf{u} =\geq (u_1, \ldots, u_D)^T$ est le vecteur D des effets aléatoires de domaine avec $\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Omega})$ où $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\sigma_u^2) = \sigma_u^2 \mathbf{I}_D$ et $\mathbf{I}_D$ est une matrice diagonale $D \times D$.

Le traitement de quasi-vraisemblance pénalisée (QVP) est une procédure d'estimation largement utilisée dans l'ajustement d'un MLGM. Il s'agit d'établir une approximation linéaire de la variable de réponse non normale, puis de poser que cette variable dépendante linéarisée est approximativement normale. La méthode QVP sert largement à l'estimation sur petits domaines, puisqu'elle est bien plus facile à appliquer dans la pratique que le meilleur prédicteur empirique. Précisons que nous employons une méthode hybride avec le traitement QVP pour l'estimation des paramètres $\boldsymbol{\beta}$ et $\mathbf{u}$ dans le MLGM (3.2) et le traitement de maximum de vraisemblance restreint pour l'estimation du paramètre de variance $\boldsymbol{\Omega}(\sigma_u^2)$. Voir à ce sujet Breslow et Clayton (1993), Saei et Chambers (2003) et Manteiga, Lombardia, Molina, Morales et Santamarìa (2007). Il est connu que, dans certains cas, le traitement QVP peut mener à des estimateurs incohérents, mais de récentes applications empiriques de la procédure QVP (Manteiga et coll., 2007), font voir que la méthode fonctionne bien dans la pratique.

Dans le modèle (3.2), les valeurs attendues de y_{si} et y_{ri} étant donné u_i viennent respectivement de $\mu_{si} = E(y_{si} \,|\, u_i) = n_i \,\text{expit}(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{z}_i^T\mathbf{u})$ et $\mu_{ri} = E(y_{ri} \mid u_i) = (N_i - n_i)\,\text{expit}(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{z}_i^T\mathbf{u})$. Avec (3.2), un prédicteur plug-in empirique (PE) du chiffre de population y_i dans le domaine i est

$$\hat{y}_i^{\text{PE}} = y_{si} + \hat{\mu}_{ri} = y_{si} + (N_i - n_i)\,\hat{\pi}_i^{\text{PE}}, \tag{3.3}$$

où $\hat{\pi}_i^{\text{PE}} = \text{expit}\left(\mathbf{x}_i^T\hat{\boldsymbol{\beta}} + \mathbf{z}_i^T\hat{\mathbf{u}}\right)$. On obtient une estimation de la proportion correspondante dans le domaine i comme $\hat{P}_i^{\text{PE}} = N_i^{-1}\hat{y}_i^{\text{PE}}$. Pour le domaine i non échantillonné avec le vecteur lié de covariables $\mathbf{x}_{i,\text{out}}$, l'estimateur synthétique de y_i dans (3.2) est $\hat{y}_i^{\text{SYN}} = N_i\hat{\pi}_i^{\text{SYN}}$, avec $\hat{\pi}_i^{\text{SYN}} = \text{expit}\left(\mathbf{x}_{i,\text{out}}^T\hat{\boldsymbol{\beta}}\right)$.

Comme il repose sur des dénombrements d'échantillon sans pondération, le modèle (3.1) implique que l'échantillonnage dans les domaines n'est pas informatif compte tenu des valeurs des variables de contexte et des effets aléatoires de domaine. C'est pourquoi le prédicteur (3.3) ne tient pas compte du plan d'échantillonnage complexe. Si le plan de sondage est informatif et qu'on dispose de chiffres pondérés

d'enquête, deux grands problèmes se posent. D'abord, les chiffres en question ne seront pas nécessairement les entiers $0, 1, \ldots, n_i$; ils prendront plutôt une valeur dans un ensemble fini de nombres en espacement inégal (non nécessairement des entiers) déterminés par la pondération d'enquête des cas de l'échantillon dans le domaine i. Ensuite, la variance d'échantillonnage estimée des chiffres pondérés d'enquête y_{si} qu'implique la distribution binomiale, $v(y_{siw}) \approx n_i p_{iw}(1 - p_{iw})$, sera erronée. Korn et Graubard (1998) proposent plutôt de modéliser l'estimation pondérée de probabilité d'enquête pour un domaine comme une proportion binomiale avec une « taille d'échantillon efficace » faisant correspondre la variance binomiale résultante à la variance d'échantillonnage réelle de l'estimation directe pondérée d'enquête du domaine. Le recours à une « taille d'échantillon efficace » a été examiné par divers auteurs, dont Mercer et coll. (2014) et Liu et coll. (2014), comme moyen d'intégration de la pondération d'enquête. Mercer et coll. (2014) font observer que le traitement en pseudovraisemblance et le traitement par taille d'échantillon efficace mènent à des estimations identiques des proportions de petits domaines. Si on utilise la taille d'échantillon efficace au lieu de la taille réelle, on se trouve à introduire la pondération d'enquête d'un échantillonnage complexe. Qui plus est, une estimation peut être d'une plus grande précision dans un échantillon complexe que dans un échantillon aléatoire simple, puisqu'on exploite mieux les données de population par un échantillon représentatif issu d'un plan d'échantillonnage approprié. Ici, nous employons l'indice (e) pour toutes les quantités à « taille d'échantillon efficace ». Nous abordons les deux problèmes évoqués en définissant une « taille d'échantillon efficace » $n_{i(e)}$, et un « chiffre d'échantillon efficace » $y_{is(e)}$, de sorte que $y_{is(e)} = n_{i(e)} p_{iw}$. Cela donne $p_{iw} = \frac{y_{is(e)}}{n_{i(e)}}$ avec son estimateur correspondant de variance $v(p_{iw})$. Nous appliquons ensuite le modèle (3.1) en posant que le chiffre d'échantillon efficace $y_{is(e)}$ dans le petit domaine i suit une distribution binomiale, c'est-à-dire que $y_{is(e)} \sim \text{Bin}\left(n_{i(e)}, \pi_i\right)$. La « taille d'échantillon efficace » $n_{i(e)}$ est donnée par $n_{i(e)} = \frac{\hat{P}_i(1-\hat{P}_i)}{v^*(p_{iw})}$, où $\hat{P}_i$ est une prédiction préliminaire par modèle de la proportion de population P_i dans un modèle linéaire généralisé et où l'estimation de variance $v^*(p_{iw})$ dépend de $\hat{P}_i$ par une fonction de variance généralisée (FVG) (voir Liu et coll. (2014)). Là, $y_{is(e)} = 0$ si $p_{iw} = 0$, mais cela ne pose aucun problème, puisque $\hat{P}_i > 0$ implique que $n_{i(e)} > 0$. À noter que nous employons une fonction de variance généralisée pour produire des estimations de la variance d'échantillonnage même dans le cas des domaines dont le chiffre de dénombrement observé est nul. De la sorte, nous n'excluons aucun domaine de l'ajustement du modèle. Nous obtenons finalement le prédicteur empirique de y_i en remplaçant (n_i, y_{si}) par $\left(n_{i(e)}, y_{is(e)}\right)$ dans (3.2), ce qui garantit que la pondération d'échantillonnage sert à l'estimation sur petits domaines. Plus précisément, le prédicteur plug-in empirique (PE) de y_i est alors

$$\hat{y}_{i(e)}^{\text{PE}} = y_{is(e)} + \left(N_i - n_{i(e)}\right)\hat{\pi}_{i(e)}^{\text{PE}}, \tag{3.4}$$

avec $\hat{\pi}_{i(e)}^{\text{PE}} = \text{expit}\left(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{(e)} + \mathbf{z}_i^T\hat{\mathbf{u}}_{(e)}\right)$. L'estimation de la proportion dans le domaine i est $\hat{P}_{i(e)}^{\text{PE}} = N_i^{-1}\hat{y}_{i(e)}^{\text{PE}}$. Dans ce cas, il s'agit respectivement avec $\hat{\boldsymbol{\beta}}_{(e)}$ et $\hat{u}_{i(e)}$ des estimations du paramètre des effets fixes et du

prédicteur du paramètre des effets aléatoires dans le modèle (3.2) compte tenu d'une « taille d'échantillon efficace » $n_{i(e)}$, et d'un « chiffre d'échantillon efficace » $y_{is(e)}$. De même, l'estimateur synthétique de y_i est $\hat{y}_{i(e)}^{\mathrm{SYN}} = N_i \hat{\pi}_{i(e)}^{\mathrm{SYN}}$, avec $\hat{\pi}_{i(e)}^{\mathrm{SYN}} = \mathrm{expit}\left(\mathbf{x}_{i,\mathrm{out}}^T \hat{\boldsymbol{\beta}}_{(e)}\right)$.

3.2 Calcul de la taille d'échantillon efficace

En nous appuyant sur les idées exposées dans Liu et coll. (2014) et Franco et Bell (2013), nous décrivons une procédure permettant de calculer par district les valeurs de taille d'échantillon efficace et le chiffre d'échantillon efficace correspondant pour le dénombrement des ménages pauvres. Nous obtenons d'abord une prédiction approximative par modèle de P_i, disons $\hat{P}_i$, à partir d'un modèle linéaire logistique ajusté à des estimations (pondérées) directes par district p_{iw} avec un ensemble de variables auxiliaires au niveau des districts. Nous ajustons ce modèle à l'aide de la fonction *glm* dans le langage R en spécifiant la famille comme « binomiale » et avec les tailles d'échantillon par district comme pondération. Par définition, les estimations par modèle $\hat{P}_i = \exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\lambda}}\right)$ se situeront dans l'intervalle (0, 1). Ici, $\mathbf{x}_i$ désigne le vecteur de covariables dans la fonction *glm* comme dans le modèle (3.1). Nous nous reportons ensuite aux estimations directes de variance $s_i^2 = v(p_{iw})$ de l'enquête du NSSO comme variables dépendantes dans un modèle FVG. Au moyen des données des districts sans dénombrement nul des ménages pauvres, nous établissons une estimation lissée de la variance d'échantillonnage par le modèle

$$E(s_i^2) = \mathrm{FVG}_i = \alpha_0 \left[P_i (1 - P_i)\right]^{\alpha 1} (R_i)^{\alpha 2}.$$

Dans ce cas, $R_i = \sum_{j=1}^{n_i} w_{ij}^2 \left\{\left(\sum_{j=1}^{n_i} w_{ij}\right)^2\right\}^{-1}$, où w_{ij} est le poids du ménage j dans le district i. Si on prend les logarithmes des deux côtés, on obtient

$$\log \mathrm{FVG}_i = \alpha_0^* + \alpha_1 \log\left[\hat{P}_i \left(1 - \hat{P}_i\right)\right] + \alpha_2 \log(R_i).$$

Ce modèle peut être ajusté avec la fonction *lm* dans R pour dégager les estimations $\hat{\alpha}_0^*$, $\hat{\alpha}_1$ et $\hat{\alpha}_2$ des coefficients de régression. Nous calculons ensuite les estimations FVG lissées de la variance d'échantillonnage comme $\hat{s}_i^2 = v^*(p_{iw}) = \exp(\hat{\alpha}_0^*) \times \left[\hat{P}_i \left(1 - \hat{P}_i\right)\right]^{\hat{\alpha}_1} \times (R_i)^{\hat{\alpha}_2}$. Enfin, nous calculons pour chaque district la taille d'échantillon efficace $n_{i(e)}$ et le chiffre d'échantillon efficace de dénombrement des ménages pauvres $y_{is(e)}$, comme $n_{i(e)} = \frac{\hat{P}_i(1-\hat{P}_i)}{v^*(p_{iw})}$ et $y_{is(e)} = n_{i(e)} p_{iw}$. Nous arrondissons les valeurs à l'entier le plus proche. La figure 3.1 présente les tailles d'échantillon efficaces par rapport aux tailles d'échantillon observées. Les chiffres d'échantillon efficaces et observés se trouvent à la figure 3.2. Dans la majorité des cas, la taille efficace est inférieure à la taille observée. De même, le chiffre d'échantillon efficace est le plus souvent inférieur au chiffre observé, indice que le plan de sondage fait perdre de l'information si on le compare pour les mêmes districts à un échantillonnage aléatoire simple. La figure 3.3 livre par district les estimations directes d'enquête, pondérées et non pondérées, de la proportion de ménages pauvres. Il ressort de cette figure que les estimations directes non pondérées sous-évaluent la proportion de ménages pauvres.

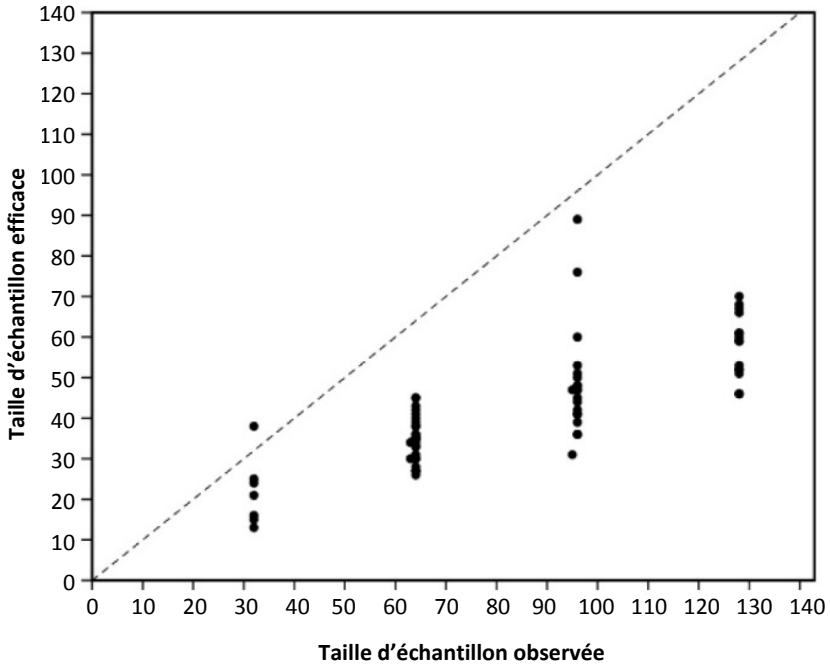

Figure 3.1 Taille d'échantillon efficace et taille d'échantillon observée pour les données du NSSO.

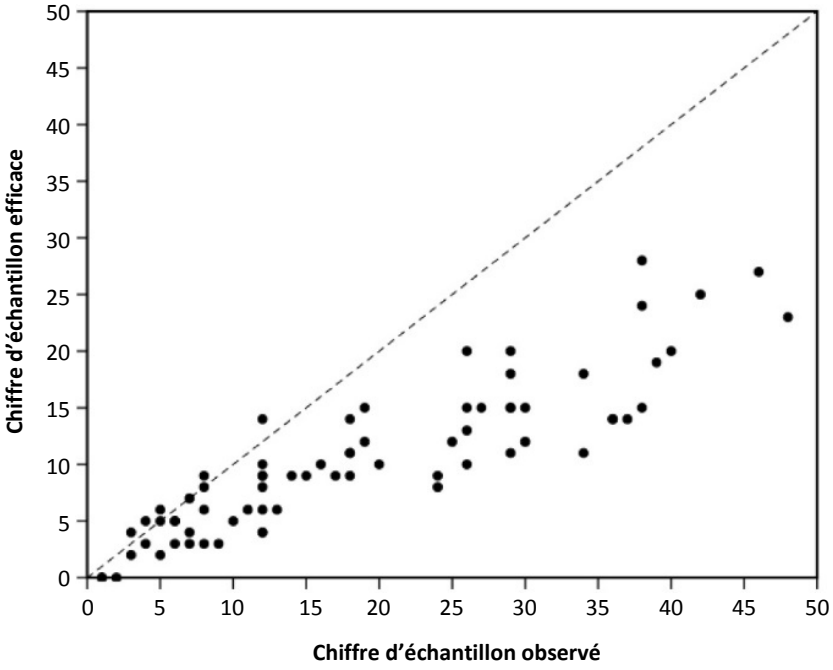

Figure 3.2 Chiffre d'échantillon efficace et chiffre d'échantillon observé pour les données du NSSO.

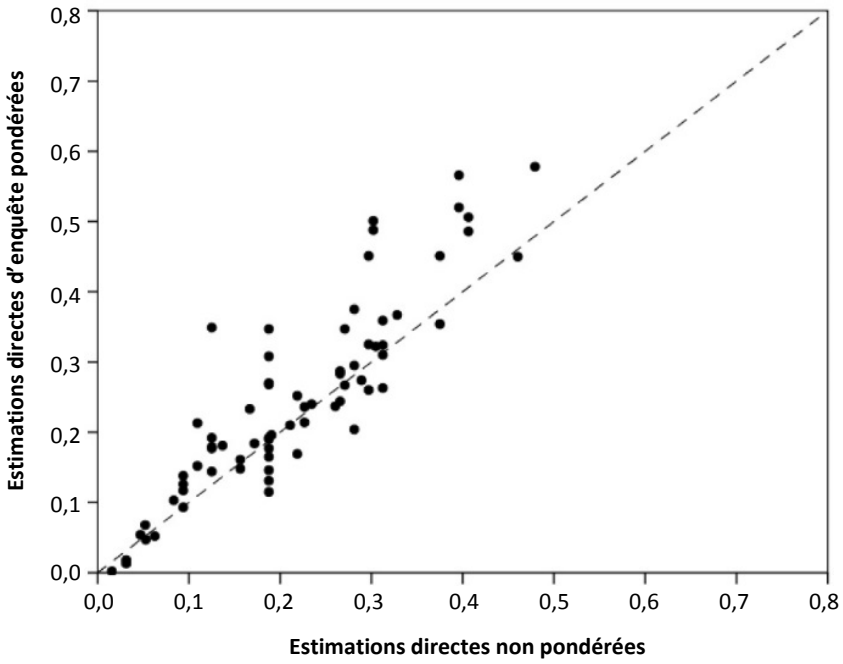

Figure 3.3 Estimations directes d'enquête pondérées et non pondérées par district de la proportion de ménages pauvres.

3.3 Effets aléatoires de domaine en corrélation spatiale

Dans le modèle (3.2), on suppose que les effets aléatoires de domaine ne sont pas en corrélation. Dans de nombreuses applications, la situation des domaines dans l'espace est le reflet d'une information de contexte manquante et l'hypothèse d'absence de corrélation est à mettre en doute dans le cas des effets aléatoires de domaine. Il est souvent raisonnable de supposer que les effets de domaines voisins (définis, par exemple, par un critère de contiguïté) sont en corrélation et que la corrélation décroît jusqu'à zéro à mesure qu'augmente la distance entre les domaines. Pour tenir compte de la dépendance entre domaines voisins, on emploie souvent dans un traitement d'EPD des modèles spatiaux pour les effets aléatoires de domaine (Pratesi et Salvati, 2008; Chandra et Salvati, 2018). Nous le faisons en introduisant une dépendance spatiale dans la structure des erreurs du modèle (3.2). Plus précisément, nous posons un processus d'autorégression simultanée (ARS) des erreurs (Pratesi et Salvati, 2008) où le vecteur des effets aléatoires de domaine $\mathbf{v} = (v_i)$ peut s'exprimer sous la forme

$$\mathbf{v} = \rho \mathbf{L}\mathbf{v} + \mathbf{u}, \tag{3.5}$$

où $\mathbf{v} = (\mathbf{I}_D - \rho\mathbf{L})^{-1}\mathbf{u}$, avec $E(\mathbf{v}) = 0$ et $\text{Var}(\mathbf{v}) = \sigma_u^2 [(\mathbf{I}_D - \rho\mathbf{L})(\mathbf{I}_D - \rho\mathbf{L}^T)]^{-1} = \mathbf{\Omega}_{sp}(\boldsymbol{\delta})$. Ici, $\mathbf{L}$ est une matrice de proximité d'ordre D, $\mathbf{u} \sim N(0, \sigma_u^2\mathbf{I}_D)$ et ρ est un coefficient d'autorégression spatiale. À partir de (3.2) et (3.5), le MLGM en dépendance spatiale (MLGMS) s'obtient par la forme

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}(\mathbf{I}_D - \rho\mathbf{L})^{-1}\mathbf{u} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}. \tag{3.6}$$

La matrice de proximité $\mathbf{L}$ décrit comment les effets aléatoires des domaines voisins sont liés, alors que ρ définit l'étroitesse de ce rapport spatial. En d'autres termes, c'est l'indicateur du degré de similitude entre un objet et les objets voisins. Dans ce qui suit, nous employons l'indice « *sp* » pour désigner les quantités liées dans le modèle (3.5). La forme la plus simple pour définir $\mathbf{L}$ est une matrice de contiguïté, c'est-à-dire une matrice carrée binaire $\mathbf{L}$ d'ordre D avec des valeurs non nulles seulement pour les paires de domaines adjacents. Pour faciliter l'interprétation, nous prêtons généralement à cette matrice une forme réduite et centrée de ligne, auquel cas ρ est appelé paramètre d'autocorrélation spatiale (Banerjee, Carlin et Gelfand, 2004). L'élément l_{jk} $(j, k = 1, \ldots, D)$ d'une matrice de contiguïté prend formellement la valeur 1 si le domaine j a une arête commune avec le domaine k et 0 dans les autres cas. Sous une forme réduite et centrée de ligne, cela devient

$$l_{jk} = \begin{cases} t_j^{-1} & \text{si } j \text{ et } k \text{ sont contigus} \\ 0 & \text{dans les autres cas,} \end{cases} \tag{3.7}$$

où t_j est le nombre total de domaines partageant une arête avec le domaine j (ce qui comprend le domaine j lui-même). La contiguïté représente sans doute la spécification la plus simple, mais non nécessairement la meilleure, d'une matrice d'interaction spatiale (voir Chandra (2013), par exemple). Il est peut-être plus informatif que cette interaction soit précisée, par exemple, comme une certaine fonction de la longueur d'une limite commune entre domaines voisins ou de la distance entre domaines. Pour notre propos, nous examinerons donc une définition fondée sur la distance pour la matrice de proximité $\mathbf{L}$ (matrice dont les entrées sont fonction de la distance entre les petits domaines ou districts). Nous nous attachons aux différentes façons de définir la proximité (ou la matrice de pondération spatiale), le but étant de reconnaître le meilleur moyen d'exploiter l'information spatiale pour produire des estimations fiables de petits domaines. Disons que la situation dans l'espace du domaine ou du district i correspond aux coordonnées d'un point spatial arbitrairement défini (coordonnées $x - y$ ou longitude et latitude dans une représentation bidimensionnelle) dans un domaine (centroïde, par exemple), ce que nous désignerons par $c_i = (c_{ix}, c_{iy})$. Soit $d_{jk} = \| c_j - c_k \|$ une mesure appropriée de la distance entre les points spatiaux des domaines j et i. Nous considérons alors les spécifications suivantes de la matrice de proximité $\mathbf{L}$ en fonction de la distance :

(i) Proximité définie comme l'inverse de la distance entre les domaines :

$$\mathbf{L} = \{l_{jk}\} = \begin{cases} d_{jk}^{-1} & j \neq k; \\ 0 & j = k. \end{cases} \tag{3.8}$$

(ii) Spécification exponentielle de la fonction de proximité ainsi définie :

$$\mathbf{L} = \{l_{jk}\} = \exp\{-0{,}5 d_{jk}^2\} \tag{3.9}$$

(iii) Spécification gaussienne de la fonction de proximité ainsi définie :

$$\mathbf{L} = \{l_{jk}\} = \exp\left\{-0{,}5 \left(d_{jk} / b\right)^2\right\}, \tag{3.10}$$

où le paramètre b est la largeur de bande, laquelle peut se définir au mieux à l'aide du critère des moindres carrés (Fotheringham, Brunsdon et Charlton, 2002). La largeur de bande est une mesure de la vitesse de décroissance de la proximité à mesure que croît la distance. Nous employons une procédure d'intervalidation pour estimer cette largeur. Dans ce cas, la proximité en (3.10) diminue exponentiellement à mesure qu'augmente la distance entre les domaines j et k. Plus particulièrement, nous employons la fonction *gwr.sel* dans le paquet *spgwr* de R pour calculer la valeur de largeur de bande. Nous utilisons également la fonction *Moran.I* du paquet *ape* de R pour vérifier la présence d'une corrélation spatiale dans les données. Selon les résultats de cette analyse, les données du NSSO présentent une certaine autocorrélation spatiale. Nous rejetons en particulier l'hypothèse nulle d'absence de corrélation spatiale au niveau de signification de 1 %.

Dans le modèle (3.5), il y a deux paramètres (σ_u^2 et ρ) à estimer. Posons que $\boldsymbol{\delta} = (\sigma_u^2, \rho)^T$. Si nous remplaçons ces paramètres inconnus par leurs valeurs estimées $\hat{\boldsymbol{\delta}} = (\hat{\sigma}_u^2, \hat{\rho})^T$, et désignons les estimateurs plug-in qui suivent par un accent circonflexe, nous définissons le prédicteur empirique spatial (PES) du chiffre de population dans le domaine i par

$$\hat{y}_{i(e)}^{\mathrm{PES}} = y_{is(e)} + \left(N_i - n_{i(e)}\right)\hat{\pi}_{i(e)}^{\mathrm{PES}}, \tag{3.11}$$

avec $\hat{\pi}_{i(e)}^{\mathrm{PES}} = \mathrm{expit}\left(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{(e)}^{sp} + \mathbf{z}_i^T\hat{\mathbf{v}}_{(e)}\right)$. Le prédicteur empirique spatial correspondant de la proportion dans le domaine i est $\hat{P}_{i(e)}^{\mathrm{PES}} = N_i^{-1}\hat{y}_{i(e)}^{\mathrm{PES}}$. Dans le cas d'un domaine non échantillonné, le prédicteur empirique spatial synthétique (désigné par SpSyn dans la suite du texte) de P_i est $\hat{P}_{i(e)}^{\mathrm{SpSyn}} = \mathrm{expit}\left(\mathbf{x}_{i,\mathrm{out}}^T\hat{\boldsymbol{\beta}}_{(e)}^{sp}\right)$. Là, $\hat{\boldsymbol{\beta}}_{(e)}^{sp}$ est l'estimation des paramètres d'effets fixes $\boldsymbol{\beta}$ et $\hat{\mathbf{v}}_{(e)}$ est la valeur prédite des effets aléatoires en corrélation spatiale $\mathbf{v}$ dans le modèle ARS avec la « taille d'échantillon efficace » et le « chiffre d'échantillon efficace ». L'estimation des paramètres de modèle inconnus en (3.6) découle de la procédure examinée à la section précédente, mais la composante de variance σ_u^2 est maintenant $\boldsymbol{\delta} = (\sigma_u^2, \rho)^T$ et l'effet aléatoire prédit $\hat{\mathbf{u}}_{(e)}$ est remplacé par $\hat{\mathbf{v}}_{(e)}$.

3.4 Estimation de l'erreur quadratique moyenne (EQM)

L'estimation d'erreur quadratique moyenne du prédicteur empirique de petits domaines (3.4) suit ce que décrivent Saei et Chambers (2003), Manteiga et coll. (2007), Johnson et coll. (2010), Chandra et coll. (2011), Chandra, Salvati et Chambers (2018) et les références présentées. Nous utilisons directement l'estimation EQM du prédicteur empirique (3.4) par l'expression (3.12) en remplaçant respectivement les tailles et les chiffres d'échantillon observés par les tailles et les chiffres efficaces comme pondération d'enquête. Dans le modèle (3.2), une estimation approximative EQM du PE (3.4) par les tailles et les chiffres d'échantillon efficaces est

$$\mathrm{eqm}\left(\hat{P}_{i(e)}^{\mathrm{PE}}\right) \approx m_{1i}\left(\hat{\sigma}_u^2\right) + m_{2i}\left(\hat{\sigma}_u^2\right) + 2m_{3i}\left(\hat{\sigma}_u^2\right). \tag{3.12}$$

Cette estimation de l'erreur quadratique moyenne est fondée sur une approximation analogue à celle du modèle linéaire mixte (voir Rao et Molina (2015), chapitre 5, pages 100-107), Saei et Chambers (2003) et Manteiga et coll. (2007)). Pour définir trois composantes de cette erreur (3.12), soit $\hat{\mathbf{\Sigma}} = \mathbf{Z}^T\mathbf{B}\mathbf{Z} + \hat{\mathbf{\Omega}}^{-1}$, $\mathbf{X}^* = \{\text{diag}\,(N_i^{-1})\}\,\mathbf{H}\mathbf{X}$ et $\mathbf{Z}^* = \{\text{diag}\,(N_i^{-1})\}\,\mathbf{H}\mathbf{Z}$, où

$$\mathbf{H} = \left.\frac{\partial h(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right|_{\hat{\boldsymbol{\eta}}=\boldsymbol{\eta}} = \text{diag}\left\{\hat{P}_{i(e)}^{\text{PE}}\left(1-\hat{P}_{i(e)}^{\text{PE}}\right);\ i = 1, \ldots, D\right\}$$

et où

$$\mathbf{B} = -\left.\frac{\partial^2 l_1}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^T}\right|_{\hat{\boldsymbol{\eta}}=\boldsymbol{\eta}} = \text{diag}\left\{n_{i(e)}\hat{P}_{i(e)}^{\text{PE}}\left(1-\hat{P}_{i(e)}^{\text{PE}}\right);\ i = 1, \ldots, D\right\}$$

est la matrice des dérivées secondes de l_1 (fonction de vraisemblance logarithmique l_1 définie par le vecteur $\mathbf{y}_{s(e)}$ étant donné $\mathbf{u}$) pour $\boldsymbol{\eta}$ à $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}$. D'après McGilchrist (1994), nous pouvons former la matrice de variances-covariances comme

$$\hat{\mathbf{V}} = \begin{bmatrix}\mathbf{X}^T\\ \mathbf{Z}^T\end{bmatrix}\mathbf{B}\begin{bmatrix}\mathbf{X} & \mathbf{Z}\end{bmatrix} + \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \hat{\mathbf{\Omega}}^{-1}\end{bmatrix} = \begin{pmatrix}\mathbf{X}^T\mathbf{B}\mathbf{X} & \mathbf{X}^T\mathbf{B}\mathbf{Z}\\ \mathbf{Z}^T\mathbf{B}\mathbf{X} & \hat{\mathbf{\Sigma}}\end{pmatrix} = \begin{pmatrix}\hat{\mathbf{V}}_{11} & \hat{\mathbf{V}}_{12}\\ \hat{\mathbf{V}}_{21} & \hat{\mathbf{V}}_{22}\end{pmatrix},$$

de sorte que

$$\hat{\mathbf{V}}^{-1} = \begin{pmatrix}\mathbf{X}^T\mathbf{B}\mathbf{X} & \mathbf{X}^T\mathbf{B}\mathbf{Z}\\ \mathbf{Z}^T\mathbf{B}\mathbf{X} & \hat{\mathbf{\Sigma}}\end{pmatrix}^{-1} = \begin{pmatrix}\hat{\mathbf{A}}_{11} & \hat{\mathbf{A}}_{12}\\ \hat{\mathbf{A}}_{21} & \hat{\mathbf{A}}_{22}\end{pmatrix}$$

où nous avons partitionné tant $\hat{\mathbf{V}}$ que son inverse $\hat{\mathbf{V}}^{-1}$ selon les dimensions de $\boldsymbol{\beta}$ et $\mathbf{u}$. Dans ce cas, $\hat{\mathbf{A}}_{11} = \left[\mathbf{X}^T\mathbf{B}\mathbf{X} - \mathbf{X}^T\mathbf{B}\mathbf{Z}\hat{\mathbf{\Sigma}}^{-1}\mathbf{Z}^T\mathbf{B}\mathbf{X}\right]^{-1}$ et $\hat{\mathbf{A}}_{22} = \hat{\mathbf{\Sigma}}^{-1} + \hat{\mathbf{\Sigma}}^{-1}\left\{(\mathbf{Z}^T\mathbf{B}\mathbf{X})\,\hat{\mathbf{A}}_{11}\,(\mathbf{X}^T\mathbf{B}\mathbf{Z})\right\}\hat{\mathbf{\Sigma}}^{-1}$. Nous posons $\Delta = \mathbf{Z}^*\hat{\mathbf{\Sigma}}^{-1}$; soit $\mathbf{Z}^*_{(k)}$ la k^{e} ligne de la matrice $\mathbf{Z}^*$, dont la dérivée est donnée par

$$\nabla_{(k)} = \left.\frac{\partial \Delta_{(k)}}{\partial \sigma_u^2}\right|_{\sigma_u^2=\hat{\sigma}_u^2} = \left.\frac{\partial\left(\mathbf{Z}^*_{(k)}\mathbf{\Sigma}^{-1}\right)}{\partial \sigma_u^2}\right|_{\sigma_u^2=\hat{\sigma}_u^2} = (\hat{\sigma}_u^2)^{-2}\,\mathbf{Z}^*_{(k)}\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Sigma}}^{-1} = \mathbf{Z}^*_{(k)}\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Omega}}^{-1}\hat{\mathbf{\Omega}}^{-1}\hat{\mathbf{\Sigma}}^{-1}$$

où $\hat{\mathbf{\Sigma}}^+ = \mathbf{Z}^T\left(\mathbf{B} + \mathbf{B}\mathbf{Z}\hat{\mathbf{\Omega}}\mathbf{Z}^T\mathbf{B}\right)\mathbf{Z}$. Dans cette notation, si nous supposons que le modèle (3.2) se vérifie et nous reportons aux tailles et aux chiffres d'échantillon efficaces, les composantes de l'estimation EQM (3.12) sont

$$\begin{aligned}
m_{1i}(\hat{\sigma}_u^2) &= \mathbf{z}_i^T\left(\mathbf{Z}^*\hat{\mathbf{\Sigma}}^{-1}\mathbf{Z}^{*T}\right)\mathbf{z}_i,\\
m_{2i}(\hat{\sigma}_u^2) &= \mathbf{z}_i^T\left(\mathbf{C}\hat{\mathbf{A}}_{11}\mathbf{C}^T\right)\mathbf{z}_i \quad \text{avec}\quad \mathbf{C} = \mathbf{X}^* - \mathbf{Z}^*\hat{\mathbf{\Sigma}}^{-1}\mathbf{Z}^T\mathbf{B}\mathbf{X},\ \text{et}\\
m_{3i}(\hat{\sigma}_u^2) &= \mathbf{z}_i^T\left\{\text{trace}\left[\left(\nabla_{(k)}\hat{\mathbf{\Sigma}}^+\nabla_{(l)}^T\right)v(\hat{\sigma}_u^2)\right]\right\}\mathbf{z}_i.
\end{aligned}$$

Là, $v(\hat{\sigma}_u^2)$ est la matrice des covariances asymptotiques de $\hat{\sigma}_u^2$, qui s'obtient comme l'inverse de la matrice d'information de Fisher appropriée pour $\hat{\sigma}_u^2$. Si nous employons l'estimation du maximum de vraisemblance restreint de σ_u^2, $v(\hat{\sigma}_u^2) = 2\left((\hat{\sigma}_u^2)^{-2}(D - 2t_1) + (\hat{\sigma}_u^2)^{-4}\,t_{11}\right)^{-1}$ avec $t_1 = (\hat{\sigma}_u^2)^{-1}\,\text{trace}\left(\hat{\mathbf{A}}_{22}\right)$ et $t_{11} = \text{trace}\left(\hat{\mathbf{A}}_{22}\hat{\mathbf{A}}_{22}\right)$.

L'estimation EQM du prédicteur empirique spatial PES (3.11) se définit de la même manière. Si nous remplaçons $\boldsymbol{\Omega}(\hat{\sigma}_u^2)$ par $\boldsymbol{\Omega}_{sp}\left(\hat{\boldsymbol{\delta}}\right)$ et $\hat{\mathbf{u}}$ par $\hat{\mathbf{v}}$ en (3.12), elle se formule ainsi :

$$\text{eqm}\left(\hat{P}_{i(e)}^{\text{PES}}\right) = m_{1i}^{sp}\left(\hat{\boldsymbol{\delta}}\right) + m_{2i}^{sp}\left(\hat{\boldsymbol{\delta}}\right) + 2m_{3i}^{sp}\left(\hat{\boldsymbol{\delta}}\right), \tag{3.13}$$

où dans (3.5) les trois composantes de (3.13) se définissent comme ci-après. Posons

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_{sp} &= \mathbf{Z}^T\mathbf{B}_{sp}\mathbf{Z} + \hat{\boldsymbol{\Omega}}_{sp}^{-1},\\
\mathbf{Z}_{sp}^{*} &= \{\text{diag}\,(N_i^{-1});\ i = 1, \ldots, D\}\,\mathbf{H}_{sp}\mathbf{Z},\\
\mathbf{X}_{sp}^{*} &= \{\text{diag}\,(N_i^{-1});\ i = 1, \ldots, D\}\,\mathbf{H}_{sp}\mathbf{X},
\end{aligned}$$

avec

$$\mathbf{B}_{sp} = \text{diag}\left\{n_d\hat{P}_{i(e)}^{\text{PES}}\left(1 - \hat{P}_{i(e)}^{\text{PES}}\right);\ i = 1, \ldots, D\right\}$$

et

$$\mathbf{H}_{sp} = \text{diag}\left\{\hat{P}_{i(e)}^{\text{PES}}\left(1 - \hat{P}_{i(e)}^{\text{PES}}\right);\ i = 1, \ldots, D\right\}.$$

Alors

$$\begin{aligned}
m_{1i}^{sp}\left(\hat{\boldsymbol{\delta}}\right) &= \mathbf{z}_i^T\left(\mathbf{Z}_{sp}^{*}\hat{\boldsymbol{\Sigma}}_{sp}^{-1}\mathbf{Z}_{sp}^{*T}\right)\mathbf{z}_i,\\
m_{2i}^{sp}\left(\hat{\boldsymbol{\delta}}\right) &= \mathbf{z}_i^T\left\{\mathbf{C}_{sp}\left(\mathbf{X}^T\mathbf{B}_{sp}\mathbf{X} - \mathbf{X}^T\mathbf{B}_{sp}\mathbf{Z}\hat{\boldsymbol{\Sigma}}_{sp}^{-1}\mathbf{Z}^T\mathbf{B}_{sp}\mathbf{X}\right)^{-1}\mathbf{C}_{sp}^T\right\}\mathbf{z}_i
\end{aligned}$$

avec $\mathbf{C}_{sp} = \mathbf{X}_{sp}^{*} - \mathbf{Z}_{sp}^{*}\hat{\boldsymbol{\Sigma}}_{sp}^{-1}\mathbf{Z}^T\mathbf{B}_{sp}\mathbf{X}$, et

$$m_{3i}^{sp}\left(\hat{\boldsymbol{\delta}}\right) = \mathbf{z}_i^T\left\{\text{trace}\left[\left(\nabla_{(k)}^{sp}\hat{\boldsymbol{\Sigma}}_{sp}^{+}\nabla_{(l)}^{spT}\right)v\left(\hat{\boldsymbol{\delta}}\right)\right]\right\}\mathbf{z}_i$$

avec $\hat{\boldsymbol{\Sigma}}_{sp}^{+} = \mathbf{Z}^T\left(\mathbf{B}_{sp} + \mathbf{B}_{sp}\mathbf{Z}\hat{\boldsymbol{\Omega}}_{sp}\mathbf{Z}^T\mathbf{B}_{sp}\right)\mathbf{Z}$. Ici, $v\left(\hat{\boldsymbol{\delta}}\right)$ est la matrice des covariances asymptotiques des estimateurs des paramètres $\hat{\boldsymbol{\delta}}$, des composantes de variance

$$\nabla_{(k)}^{sp} = \left.\frac{\partial\Delta_{(k)}^{sp}}{\partial\hat{\boldsymbol{\delta}}}\right|_{\hat{\boldsymbol{\delta}}=\boldsymbol{\delta}} = \left.\frac{\partial\left(\mathbf{Z}_{"sp(k)}^{*}\boldsymbol{\Sigma}_{sp}^{-1}\right)}{\partial\left(\sigma_u^2,\ \rho\right)^T}\right|_{\hat{\boldsymbol{\delta}}=\boldsymbol{\delta}} = \mathbf{Z}_{sp(k)}^{*}\hat{\boldsymbol{\Sigma}}_{sp}^{-1}\boldsymbol{\Omega}_{sp}^{-1}\left(\hat{\boldsymbol{\delta}}\right)\boldsymbol{\Omega}_{sp}^{-1}\left(\hat{\boldsymbol{\delta}}\right)\hat{\boldsymbol{\Sigma}}_{sp}^{-1},$$

où $\Delta^{sp} = \mathbf{Z}_{sp}^{*}\hat{\boldsymbol{\Sigma}}_{sp}^{-1}$ and $\mathbf{Z}_{sp(k)}^{*}$ est la k^{e} ligne de $\mathbf{Z}_{sp}^{*}$.

4 Résultats empiriques

Dans cette section, nous examinerons l'application des méthodes d'EPD présentées à la section précédente à des données réelles d'enquête pour produire des estimations de la proportion de ménages ruraux pauvres au niveau des districts de l'État d'Uttar Pradesh en Inde. Nous exploitons les données de l'enquête 2011-2012 sur les dépenses de consommation des ménages du NSSO et les données du recensement 2011 de la population en Inde et posons une spécification binomiale pour les chiffres d'échantillon « efficaces » de dénombrement des ménages ruraux pauvres au niveau des districts. Nous récapitulons au tableau 4.1 les prédicteurs de petits domaines que nous retenons et les estimateurs EQM qui y sont liés. Nous décrivons d'abord un certain nombre d'importantes mesures de diagnostic pouvant servir

à l'examen des hypothèses des modèles sous-jacents et à la validation des résultats empiriques des différentes méthodes d'EPD.

Tableau 4.1
Définition des divers prédicteurs de petits domaines

Prédicteur	Description	Estimation EQM
DIR	Estimation directe d'enquête	Variance de DIR à la section 3
PE	PE (3.4) dans le modèle (3.1)	Estimation EQM (3.12)
PES1	PE spatial (3.11) dans le modèle (3.6) + pondération (3.7)	Estimation EQM (3.13)
PES2	PE spatial (3.11) dans le modèle (3.6) + pondération (3.8)	Estimation EQM (3.13)
PES3	PE spatial (3.11) dans le modèle (3.6) + pondération (3.9)	Estimation EQM (3.13)
PES4	PE spatial (3.11) dans le modèle (3.6) + pondération (3.10)	Estimation EQM (3.13)

4.1 Mesures de diagnostic

Dans les applications d'EPD, deux types de mesures de diagnostic sont couramment utilisés, à savoir les diagnostics de modèle et les diagnostics d'estimations sur petits domaines (Brown, Chambers, Heady et Heasman, 2001). Les diagnostics de modèle servent avant tout à vérifier les hypothèses de distribution du modèle de petits domaines sous-jacent, c'est-à-dire à déterminer le rendement de ce modèle de travail en ajustement aux données d'enquête. Les diagnostics d'estimations sur petits domaines livrent une indication de la fiabilité (et de la validité) des estimations de modélisation produites par les différentes méthodes d'EPD.

Dans les modèles de petits domaines définis en (3.1) et (3.6), on suppose que les effets aléatoires de domaine sont d'une distribution normale avec moyenne nulle et variance fixe. Si les hypothèses de modélisation se vérifient, on s'attend à ce que les résidus au niveau des domaines (ou districts) soient d'une distribution aléatoire autour de zéro. Ces résidus se calculent respectivement comme $r_i = \hat{\eta}_{i(e)} - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{(e)}$ et $r_i = \hat{\eta}_{i(e)} - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{(e)}^{sp}$ dans les modèles (3.1) et (3.6). On peut se reporter à des histogrammes et à des courbes normales de probabilité $(\mathrm{q} - \mathrm{q})$ pour examiner l'hypothèse de normalité. Les figure 4.1a, b et c présentent l'histogramme des résidus au niveau des districts (graphique du haut), la courbe normale de probabilité $(\mathrm{q} - \mathrm{q})$ des résidus à ce même niveau (graphique du milieu) et la distribution des résidus dans les districts (graphique du bas) selon les méthodes de petits domaines. Nous employons également le test de Shapiro-Wilk (fonction *shapiro.test* dans R) pour vérifier la normalité des effets aléatoires de domaine. Le tableau 4.2 présente les résultats de ce test pour les diverses méthodes d'EPD avec des valeurs p de moins de 0,05 indiquant que les données s'écartent de la normalité. Dans les figures 4.1a, b et c, les résidus au niveau des districts pour toutes les méthodes d'EPD semblent d'une distribution aléatoire autour de zéro, comme on pouvait s'y attendre. De plus, les histogrammes et les courbes $\mathrm{q} - \mathrm{q}$ aux figures 4.1a, b et c accréditent cette même hypothèse de normalité. Au tableau 4.2, nous pouvons aussi voir que les valeurs p du test de Shapiro-Wilk sont élevées pour toutes les méthodes d'EPD. Nous en concluons que les effets aléatoires de domaine sont sans doute d'une distribution normale.

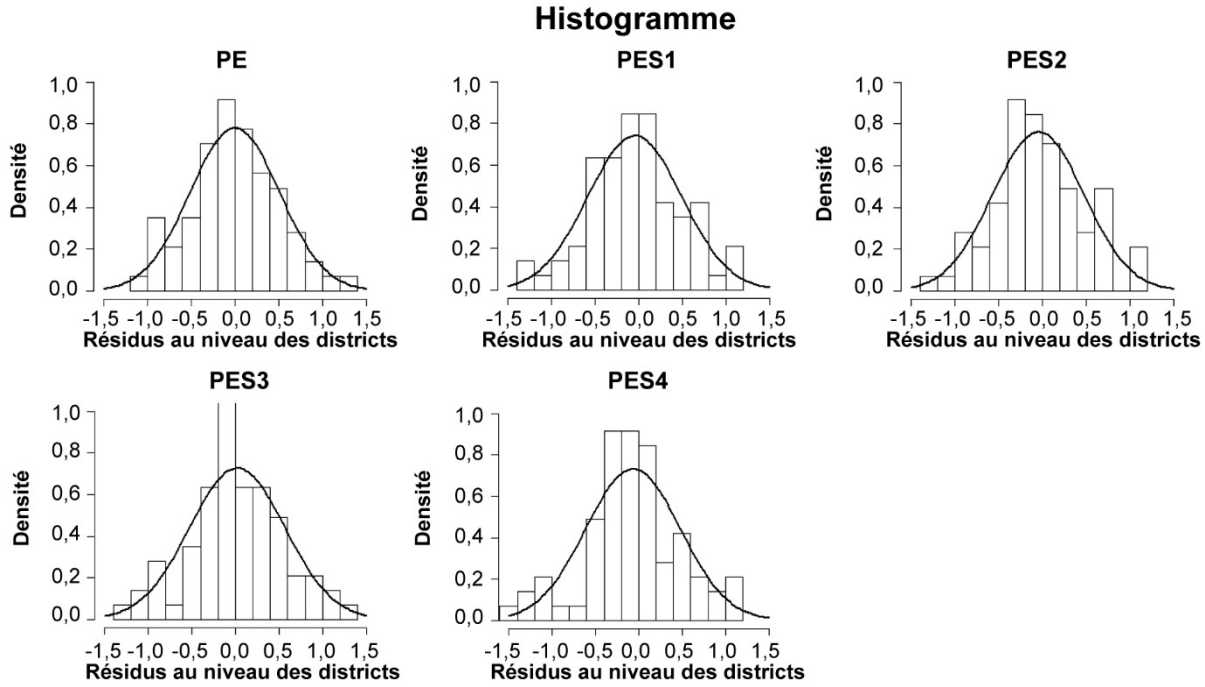

Figure 4.1a Histogrammes des résidus au niveau des districts selon les méthodes d'EPD.

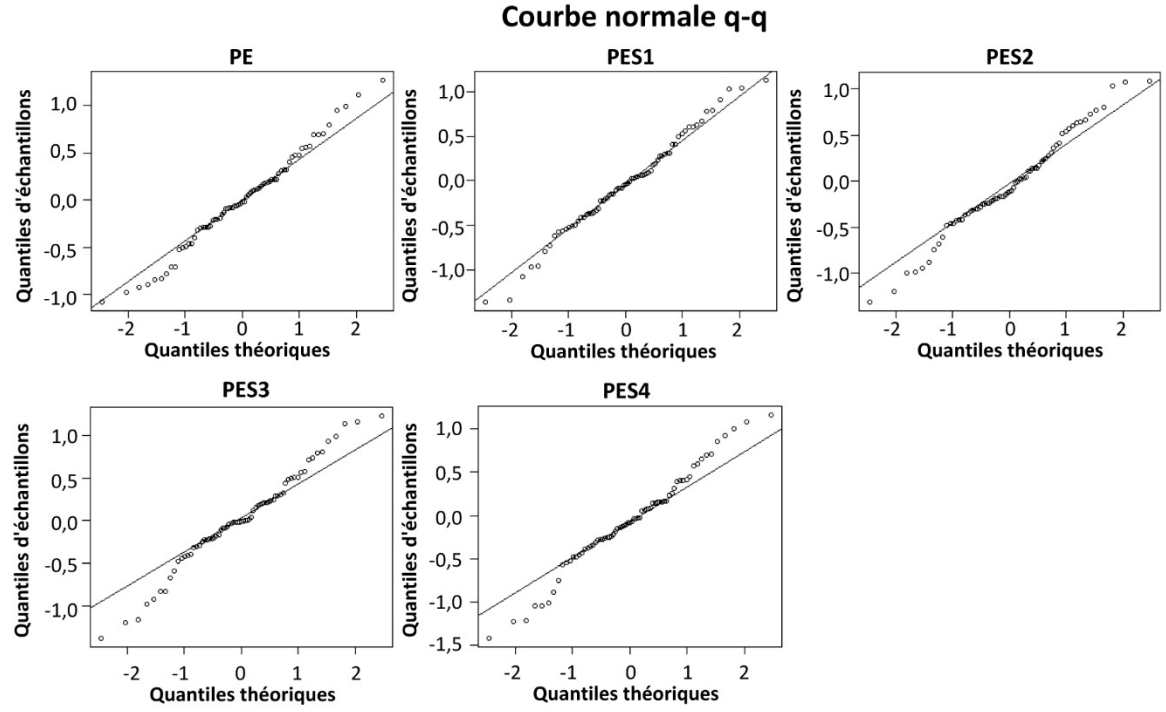

Figure 4.1b Courbes normales (q − q) des résidus au niveau des districts selon les méthodes d'EPD.

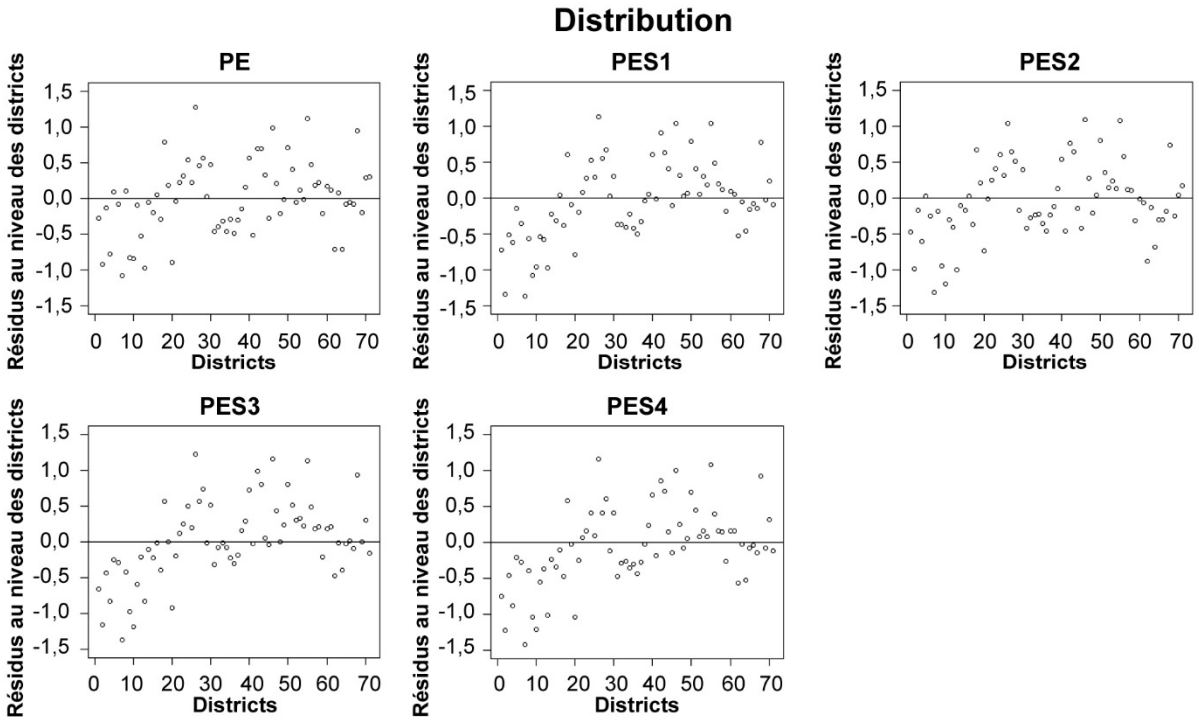

Figure 4.1c Distributions des résidus au niveau des districts selon les méthodes d'EPD.

Tableau 4.2
Résultats du test de Shapiro-Wilk de la normalité des effets aléatoires au niveau des districts

Méthode d'EPD	Valeur de statistique de test (W)	Valeur p
PE	0,988	0,764
PES1	0,989	0,771
PES2	0,984	0,491
PES3	0,982	0,391
PES4	0,982	0,399

Brown et coll. (2001) décrivent un ensemble de diagnostics pouvant permettre de jauger la validité et la fiabilité des estimations sur petits domaines par modèle. Ceux-ci reposent sur l'argument selon lequel de telles estimations devraient être a) cohérentes par rapport aux estimations directes d'enquête non biaisées (elles devraient approcher les estimations directes d'enquête en toute cohérence, c'est-à-dire en étant proches des valeurs attendues de ces dernières) et b) plus précises que les estimations directes d'enquête comme l'indiqueraient des estimations moindres d'erreur quadratique moyenne (les estimations sur petits domaines par modèle devraient présenter des EQM significativement moindres que les variances des estimations directes d'enquête correspondantes). Nous allons considérer quatre mesures de diagnostic communément utilisées qui satisfont à ces exigences, à savoir un diagnostic de biais, un test de qualité d'ajustement, un diagnostic par intervalle de confiance à 95 % et enfin un diagnostic de coefficient de variation (CV) en pourcentage. Les deux premiers portent sur la validité et les deux autres, sur la fiabilité ou le gain de précision des estimations sur petits domaines par modèle. Nous utilisons en outre un diagnostic de calage où les estimations par modèle sont agrégées à un niveau supérieur et comparées à ce niveau aux

estimations directes d'enquête (voir, par exemple, Chandra et coll. (2011). À noter que, dans ce cas, les estimations directes se définissent comme estimations directes pondérées d'enquête.

Le diagnostic de biais repose sur l'idée que les estimations directes d'enquête sont des estimations sans biais des valeurs d'intérêt de la population (valeurs réelles). Si les estimations sur petits domaines par modèle sont « proches » des valeurs d'intérêt, les estimations directes d'enquête sans biais devraient se comporter comme des variables aléatoires dont les valeurs attendues correspondraient aux valeurs des estimations par modèle. Il est donc possible d'établir un diagnostic de qualité d'ajustement qui équivaut au test de Wald, le but étant de juger si les différences entre les estimations directes et les estimations par modèle ont une moyenne nulle (Brown et coll., 2001). La statistique du test de Wald se calcule comme

$$\mathrm{W} = \sum_i \frac{\left(\mathrm{DIR}_i - \text{Estimation par modèle}_i\right)^2}{v\left(\mathrm{DIR}_i\right) + \mathrm{eqm}\left(\text{Estimation par modèle}_i\right)}.$$

La valeur de W est comparée au 95e percentile d'une distribution khi-carré à D degrés de liberté. Dans ce cas, $D = 71$ et la valeur de ce 95e percentile est 91,670. Le tableau 4.3 présente les résultats du diagnostic de qualité d'ajustement. Nous avons également calculé le biais moyen (Biais) et la différence relative moyenne (RE) entre les estimations directes et les estimations par modèle :

$$\text{Biais} = D^{-1}\left(\sum_i \text{Estimation par modèle}_i - \mathrm{DIR}_i\right)$$

et

$$\mathrm{RE} = D^{-1}\sum_i\left(\frac{\text{Estimation par modèle}_i - \mathrm{DIR}_i}{\mathrm{DIR}_i}\right).$$

Le tableau 4.4 indique les valeurs de Biais et RE pour différentes méthodes d'EPD. Les résultats aux tableaux 4.3 et 4.4 montrent clairement que les estimations par modèle issues des différentes méthodes d'EPD sont cohérentes par rapport aux estimations directes d'enquête. Enfin, nous présentons à la figure 4.2 un ensemble de courbes de diagnostic de biais où nous rapprochons les estimations directes d'enquête (axe des Y) des estimations par modèle correspondantes (axe des X) et vérifions la divergence entre la courbe de régression et la ligne $Y = X$. Ces tracés indiquent que les estimations par modèle sont moins extrêmes que les estimations directes d'enquête, ce qui illustre le résultat caractéristique du traitement d'EPD qui rabat les valeurs plus extrêmes vers la moyenne. Les valeurs de R^2 pour la courbe de régression ajustée (par les moindres carrés ordinaires) entre les estimations directes d'enquête et les estimations PE, PES1, PES2, PES3 et PES4 sont respectivement de 97, 93, 94, 92 et 96 %. En règle générale, ces différents diagnostics de biais font tous voir que les estimations issues de toutes les méthodes d'EPD seraient cohérentes par rapport aux estimations directes d'enquête.

Tableau 4.3
Diagnostic de qualité d'ajustement

Méthode d'EPD	Qualité d'ajustement*
PE	23,645
PES1	30,727
PES2	27,784
PES3	30,930
PES4	24,442

* Une faible valeur (< 91,670 dans ce cas) indique l'absence de différence statistiquement significative entre les estimations par modèle et les estimations directes.

Tableau 4.4
Valeurs de diagnostic de biais entre les estimations directes (pondérées) d'enquête et les estimations par modèle

Méthode d'EPD	Biais	RE
PE	0,0023	0,2068
PES1	0,0007	0,2155
PES2	0,0016	0,1935
PES3	0,0013	0,2096
PES4	0,0016	0,2028

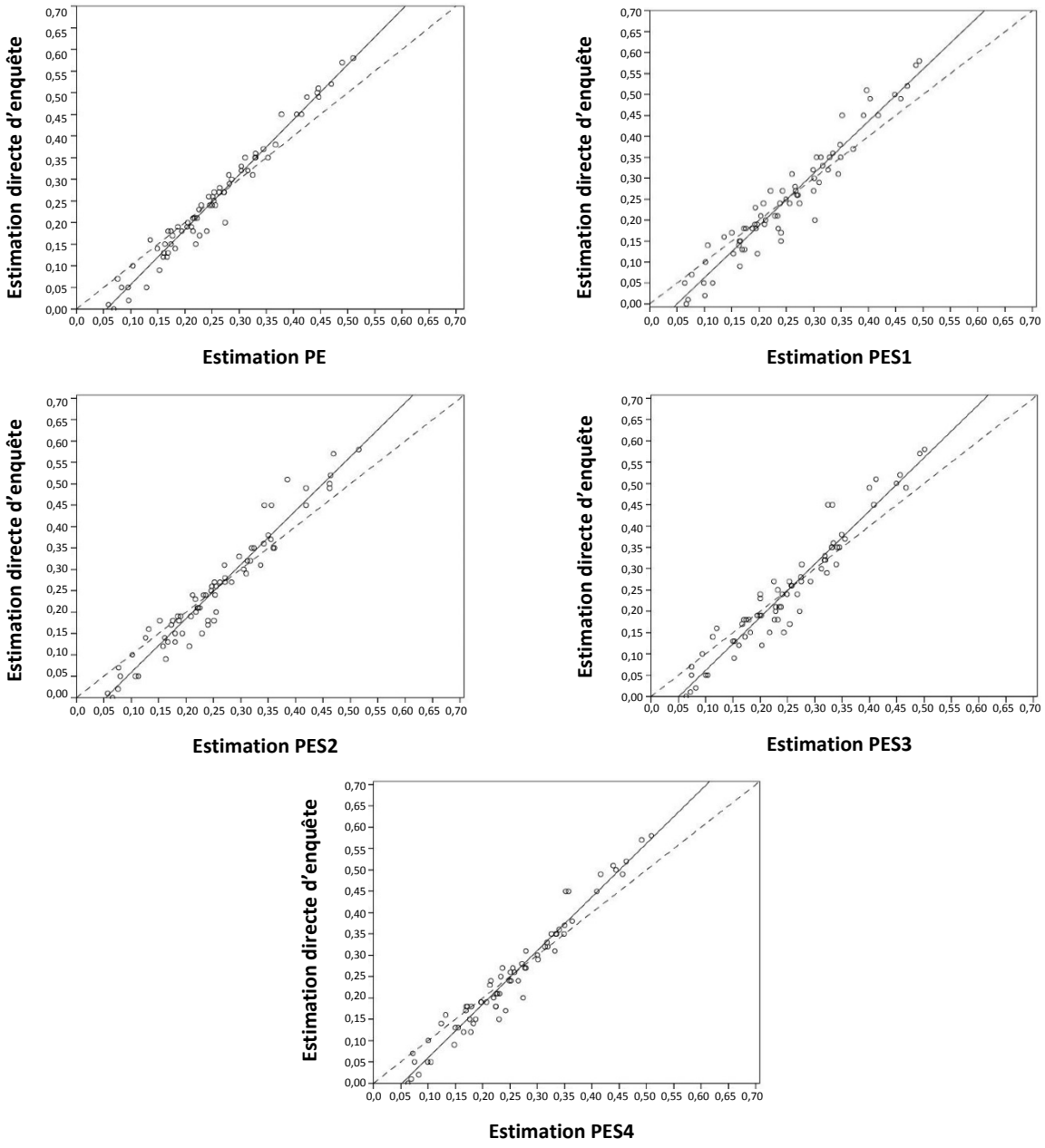

Figure 4.2 Courbes de diagnostic de biais avec la ligne $y = x$ (en pointillé) et la courbe de régression (trait uni) pour la proportion de ménages ruraux pauvres dans l'État d'Uttar Pradesh : estimations par modèle et estimations directes (pondérées) d'enquête.

Un second ensemble de diagnostics permet d'évaluer la fiabilité et le gain de précision des estimations par modèle par rapport aux estimations directes d'enquête. Le coefficient de variation (CV) en pourcentage est l'erreur-type d'échantillonnage estimée en proportion de l'estimation. Nous considérons que les estimations sur petits domaines dont les CV sont élevés ne sont pas fiables. Aucune norme internationale ne nous dit dans ce contexte ce que serait un CV « trop élevé » (Chandra et coll. (2011) et Johnson et coll. (2010)), mais en règle générale, les CV de moins de 10 % sont préférables dans le cas des estimations au niveau des districts. Le tableau 4.5 récapitule les CV des estimations directes d'enquête et des estimations par modèle issues des différentes méthodes d'EPD. Les résultats au tableau 4.5 autorisent deux conclusions. D'abord, les CV des estimations directes sont supérieurs à ceux des estimations par modèle. De plus, le rendement relatif des méthodes par modèle par rapport aux estimations directes d'enquête s'améliore à mesure que décroît la taille d'échantillon observée (ou efficace) par district. En d'autres termes, les estimations calculées par les différents modèles sont plus sûres et plus révélatrices de la fréquence de la pauvreté en milieu rural dans l'État d'Uttar Pradesh. En second lieu, si nous considérons les cinq méthodes d'EPD par modèle visées dans cette analyse, nous constatons que les CV du PE spatial (3.11) dans le modèle spatial (3.6) et avec la matrice de proximité définie par les quatre fonctions de proximité spatiale correspondant à PES1, PES2, PES3 et PES4 sont tous inférieurs au CV du prédicteur empirique (3.4) dans le modèle non spatial (3.1). En d'autres termes, le recours à l'information spatiale améliore le rendement de la méthode d'EPD. En particulier, PES3 serait la méthode la plus performante des quatre méthodes spatiales que nous avons examinées. Dans ce qui suit, nous nous attacherons donc au seul prédicteur PES3.

Tableau 4.5
Distributions des CV selon les méthodes

Valeur	DIR	PE	PES1	PES2	PES3	PES4
Minimum	12,92	13,02	12,82	12,73	12,62	12,73
1er quartile	22,16	20,26	19,02	19,91	19,02	19,59
Médiane	30,81	24,42	23,76	24,20	22,92	23,78
Moyenne	35,68	25,35	24,11	24,94	23,95	24,80
3e quartile	47,64	30,52	27,78	29,12	27,97	29,36
Maximum	99,06	43,21	39,49	42,97	38,56	42,00

La figure 4.3 présente la distribution par district des CV des méthodes DIR, PE et PES3. On peut voir que les estimations directes d'enquête de la fréquence de la pauvreté sont instables avec des CV variant de 12,92 à 99,06 % pour une moyenne de 35,68 %. Ajoutons que les CV de ces estimations sont de plus de 40 % et 50 % respectivement dans 7 et 14 des districts sur 71 (voir la figure 4.3). Par contraste, les CV moyens de PE et PES3 sont respectivement de 25,35 % et 23,95 % et le CV de PES3 est inférieur au CV de PE dans 69 districts sur 71.

La figure 4.4 trace par district la courbe des intervalles de confiance (IC) à 95 % pour DIR et PES3. Dans ce cas, les IC à 95 % sont plus larges pour les estimations directes que pour les estimations par modèle dans PES3. Nous complétons cet examen à vue des IC à 95 % en calculant les valeurs du diagnostic de recouvrement décrit dans Brown et coll. (2001) comme moyen d'évaluation de la validité des intervalles de confiance des estimations par modèle. Nous calculons d'abord des intervalles de confiance à 95 % corrigés pour les estimations directes et les estimations par modèle à l'aide des valeurs critiques par district

$$c_i = 1{,}96\left\{\sqrt{v(\mathrm{DIR}_i) + \mathrm{eqm}(\text{estimation par modèle}_i)}\right/\left(\sqrt{v(\mathrm{DIR}_i)} + \sqrt{\mathrm{eqm}(\text{estimation par modèle}_i)}\right)\}.$$

Nous comptons ensuite les fois que les intervalles corrigés ne sont pas en recouvrement. En principe, le taux de non-recouvrement devrait être d'au plus 5 %. Les taux observés de non-recouvrement sont respectivement de 1,41 % et 2,82 % pour les méthodes PE et PES3.

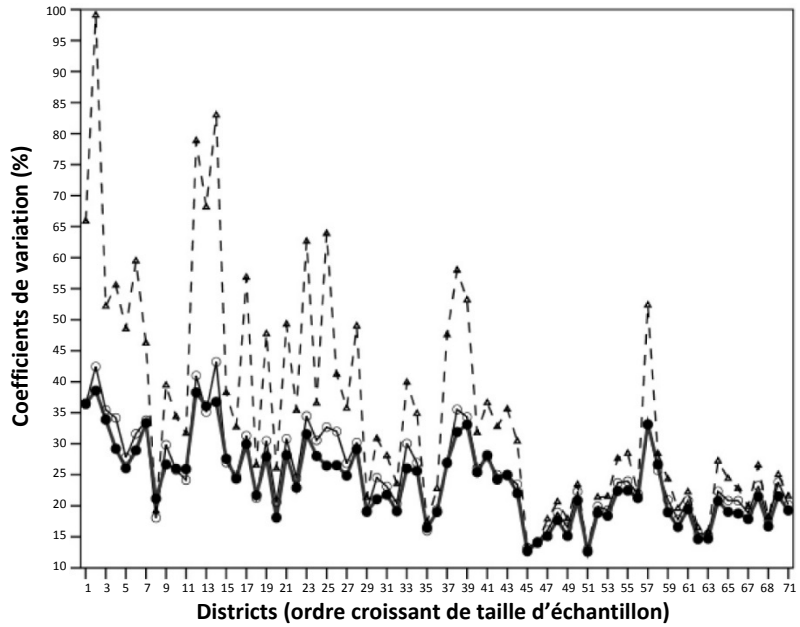

Figure 4.3 Courbe des coefficients de variation (en pourcentage) par district pour les estimations DIR (tireté, Δ), PE (trait épais,°) et PES3 (trait uni,•).

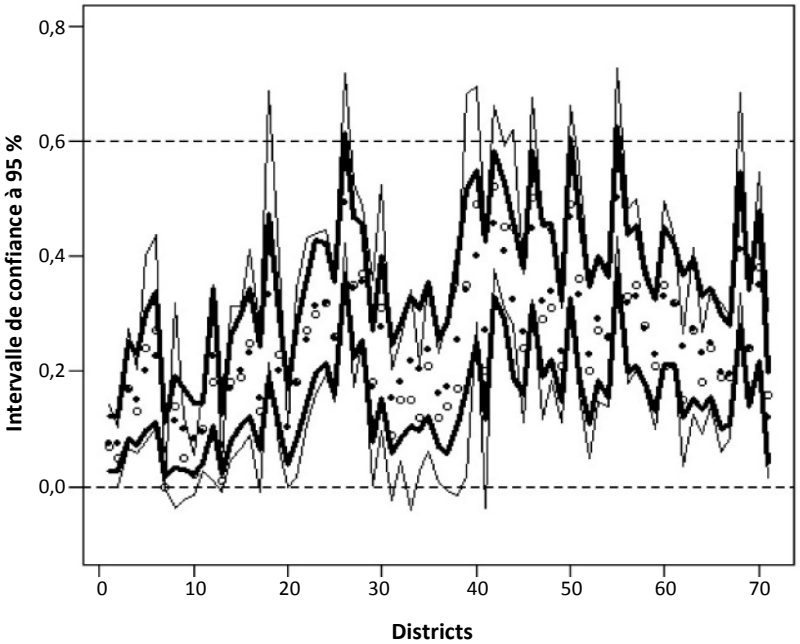

Figure 4.4 Courbe des intervalles de confiance à 95 % par district pour les estimations directes (trait épais, °) par rapport aux estimations PES3 (trait uni,•).

Enfin, nous examinons les propriétés de calage des estimations par modèle au niveau des districts en portant le tout à un plus haut niveau (État ou région, par exemple). Soit $\hat{P}_i$ et N_i l'estimation de la proportion de ménages pauvres et la taille de la population dans le district i. L'estimation au niveau de l'État de la proportion de ménages pauvres se calcule comme $\hat{P} = \sum_{i=1}^{D} N_i \hat{P}_i \Big/ \sum_{i=1}^{D} N_i$. L'État d'Uttar Pradesh se divise en régions du centre, de l'est, de l'ouest et du sud, et il est possible de regarder les propriétés de calage en fonction de ces régions. Le tableau 4.6 présente au niveau de l'État d'Uttar Pradesh et de ses régions la proportion de ménages ruraux pauvres selon les différentes méthodes d'EPD. Si nous comparons ces estimations aux estimations directes correspondantes, nous constatons que les estimations par modèle sont très proches de ces dernières au niveau de l'État et au niveau individuel pour ces régions.

Tableau 4.6
Estimation à un niveau d'agrégation de la proportion de ménages pauvres selon les différentes méthodes d'EPD. Les estimations sont agrégées sur 71 districts au niveau de l'État ainsi qu'au niveau des quatre régions

Région	DIR	PE	PES1	PES2	PES	PES4
État	0,26	0,26	0,26	0,26	0,26	0,26
Centre	0,34	0,32	0,34	0,33	0,33	0,33
Est	0,31	0,30	0,31	0,30	0,31	0,31
Sud	0,26	0,27	0,27	0,27	0,27	0,28
Ouest	0,15	0,17	0,16	0,17	0,16	0,16

4.2 Analyse des résultats

Les éléments d'analyse de la sous-section précédente démontrent clairement que les estimations par modèle issues des diverses méthodes de notre étude sont cohérentes par rapport aux estimations directes d'enquête. De plus, les valeurs des coefficients de variation, les intervalles de confiance à 95 % et les valeurs de diagnostic de recouvrement indiquent que les estimations par modèle sont fiables et plus stables que les estimations directes d'enquête correspondantes. Il est donc possible de recommander les EPD par modèle et en particulier les estimations du prédicteur PES3 aux divers intervenants s'occupant de planification et d'application des politiques. À la figure 4.5, nous présentons une carte de la pauvreté avec les proportions estimées de ménages ruraux pauvres (fréquence de la pauvreté) dans les différents districts de l'État d'Uttar Pradesh selon la méthode PES3. Cette carte décrit la distribution de la pauvreté dans cet État et fait voir le degré d'inégalité entre les districts dans la distribution des ménages ruraux pauvres, le ménage pauvre se définissant comme ayant des dépenses mensuelles de consommation par personne qui se situent sous la ligne de pauvreté de l'État. La figure 4.6 offre en pendant une carte des CV des estimations PES3 dans les différents districts. Ces cartes sont complétées par les résultats du tableau 4.7 où figurent les estimations par district avec les CV et les intervalles de confiance à 95 % pour les prédicteurs DIR et PES3. À titre d'exemple, on peut constater que, dans la région de l'ouest de l'État d'Uttar Pradesh, nombreux sont les districts à faible fréquence de la pauvreté rurale : Saharanpur, Hathras, Meerut, Baghpat, Muzaffarnagar, Bulandshahar, etc. De même, on peut voir que, dans la région de l'est et celle de Bundelkhand (partie nord-est de la carte), un certain nombre de districts (Azamgarh, Sitapur, Chitrakoot, Bahraich, Siddharthnagar, Banda, Fatehpur, Basti, Kaushambi, etc.) se caractérisent, eux, par une haute fréquence de la pauvreté. Au tableau 4.6, la fréquence de la pauvreté rurale estimée par le prédicteur empirique spatial est la plus basse dans la région de l'ouest (16%) et la plus haute dans la région du centre (33%) et la moyenne de l'État est d'environ 26 %. Au tableau 4.7, les estimations de fréquence par district pour le prédicteur PES3 varient

entre un minimum de 6 % et un maximum de 50 %. La figure 4.5 fait nettement ressortir l'inégalité spatiale de la répartition de la pauvreté rurale entre les districts de l'État d'Uttar Pradesh, tout comme les estimations des tableaux 4.6 et 4.7. Ces indications devraient se révéler utiles aux planificateurs et aux administrateurs des politiques qui désirent prendre des décisions financières et administratives efficaces.

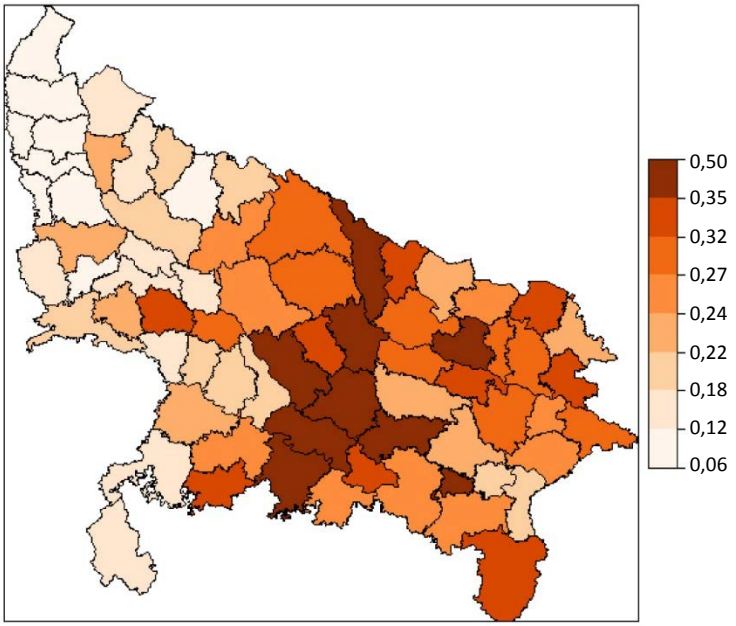

Figure 4.5 Distribution au niveau des districts de la fréquence de la pauvreté rurale dans l'État d'Uttar Pradesh selon la méthode PES3.

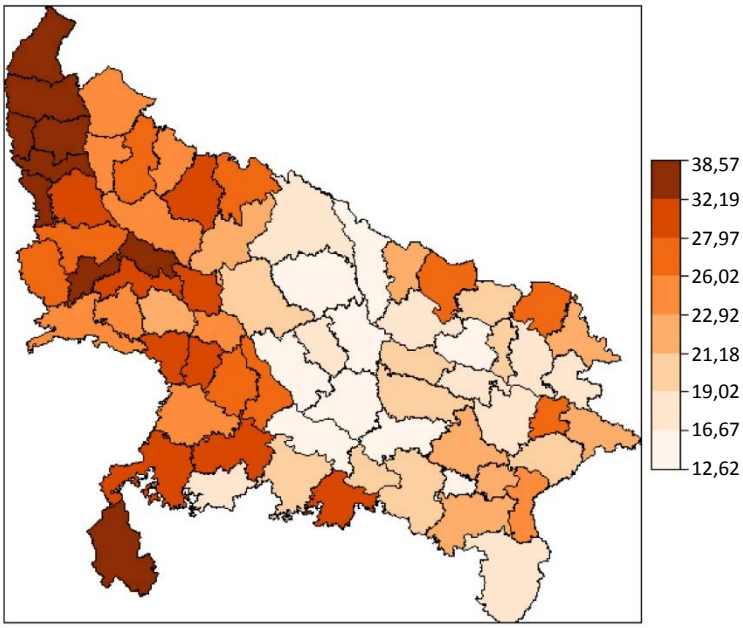

Figure 4.6 Distribution au niveau des districts des coefficients de variation en pourcentage pour la méthode PES3.

Tableau 4.7
Estimations directes (DIR) et par modèle (PES3) au niveau des districts avec leurs CV (en pourcentage) et les intervalles de confiance à 95 % correspondants pour la fréquence de la pauvreté dans l'État d'Uttar Pradesh en 2011-2012

Région	District	Fréquence estimée de la pauvreté		Intervalles de confiance à 95 %				Coefficient de variation en %	
		DIR	PES3	DIR		PES3		DIR	PES3
				Min.	Max.	Min.	Max.		
Ouest	Saharanpur	0,07	0,07	0,00	0,14	0,03	0,12	53,16	33,10
	Muzaffarnagar	0,05	0,07	0,00	0,10	0,03	0,12	52,36	33,10
	Bijnor	0,17	0,17	0,06	0,28	0,08	0,25	31,79	25,41
	Moradabad	0,13	0,15	0,06	0,20	0,07	0,23	28,41	26,67
	Rampur	0,24	0,20	0,08	0,40	0,10	0,30	34,35	25,98
	Jyotiba P, Nagar	0,27	0,23	0,10	0,44	0,11	0,34	31,69	25,92
	Meerut	0,00	0,06	0,00	0,00	0,02	0,11	78,83	38,27
	Baghpat	0,14	0,11	-0,04	0,32	0,03	0,19	65,87	36,49
	Ghaziabad	0,05	0,10	-0,02	0,12	0,03	0,17	68,16	36,06
	G, B, Nagar	0,02	0,08	-0,02	0,06	0,02	0,14	99,06	38,56
	Bulandshahr	0,10	0,09	0,03	0,17	0,04	0,15	36,60	28,15
	Aligarh	0,18	0,23	0,01	0,35	0,11	0,35	47,56	26,91
	Mahamaya Nr	0,01	0,07	-0,01	0,03	0,02	0,12	82,97	36,75
	Mathura	0,18	0,17	0,05	0,31	0,08	0,26	38,23	27,59
	Agra	0,19	0,20	0,07	0,31	0,11	0,30	32,75	24,25
	Firozabad	0,25	0,23	0,09	0,41	0,12	0,34	32,69	24,38
	Etah	0,13	0,15	-0,01	0,27	0,06	0,24	56,82	29,95
	Mainpuri	0,45	0,33	0,22	0,68	0,19	0,47	26,57	21,72
	Budaun	0,23	0,20	0,07	0,39	0,10	0,30	35,57	25,00
	Bareilly	0,05	0,10	-0,01	0,11	0,04	0,17	57,98	31,89
	Pilibhit	0,18	0,18	0,01	0,35	0,08	0,28	47,72	27,93
	Shahjahanpur	0,27	0,25	0,11	0,43	0,14	0,36	30,39	22,01
	Farrukhabad	0,18	0,17	0,01	0,35	0,08	0,27	49,26	28,16
	Kannauj	0,31	0,28	0,10	0,52	0,15	0,40	35,37	22,92
	Etawah	0,09	0,15	-0,02	0,20	0,06	0,25	62,63	31,55
	Auraiya	0,15	0,18	0,04	0,26	0,08	0,28	36,53	28,02
	Kansiram Nr	0,16	0,12	0,02	0,30	0,04	0,20	46,20	33,33
Centre	Kheri	0,30	0,31	0,16	0,44	0,20	0,43	24,31	18,96
	Sitapur	0,32	0,32	0,20	0,44	0,22	0,42	19,56	16,59
	Hardoi	0,26	0,26	0,15	0,37	0,16	0,36	22,25	19,46
	Unnao	0,57	0,49	0,42	0,72	0,37	0,61	13,36	12,69
	Lucknow	0,35	0,35	0,17	0,53	0,22	0,47	26,02	18,10
	Rae Bareli	0,37	0,36	0,25	0,49	0,25	0,46	16,46	14,64
	Kanpur Dehat	0,15	0,22	-0,04	0,34	0,10	0,33	63,88	26,47
	Kanpur Nagar	0,12	0,20	0,02	0,22	0,10	0,31	41,19	26,53
	Fatehpur	0,52	0,46	0,38	0,66	0,33	0,58	13,88	14,04
Sud	Jalaun	0,21	0,24	0,06	0,36	0,12	0,35	35,70	24,86
	Jhansi	0,12	0,16	0,00	0,24	0,07	0,25	48,97	29,13
	Lalitpur	0,14	0,17	0,00	0,28	0,06	0,29	52,16	33,90
	Hamirpur	0,17	0,25	-0,02	0,36	0,11	0,40	55,55	29,20
	Banda	0,49	0,40	0,28	0,70	0,25	0,55	21,36	19,04
	Chitrakoot	0,20	0,27	-0,03	0,43	0,12	0,43	59,47	28,95
	Mahoba	0,35	0,33	0,20	0,50	0,21	0,45	21,38	18,87

Tableau 4.7 (suite)
Estimations directes (DIR) et par modèle (PES3) au niveau des districts avec leurs CV (en pourcentage) et les intervalles de confiance à 95 % correspondants pour la fréquence de la pauvreté dans l'État d'Uttar Pradesh en 2011-2012

Région	District	Fréquence estimée de la pauvreté		Intervalles de confiance à 95 %				Coefficient de variation en %	
		DIR	PES3	DIR		PES3		DIR	PES3
				Min.	Max.	Min.	Max.		
Est	Maharajganj	0,35	0,34	0,02	0,68	0,17	0,52	48,51	26,07
	Pratapgarh	0,45	0,41	0,31	0,59	0,29	0,53	15,77	14,71
	Kaushambi	0,45	0,32	0,28	0,62	0,19	0,46	19,16	21,16
	Allahabad	0,24	0,27	0,11	0,37	0,16	0,38	27,21	20,78
	Bara Banki	0,50	0,45	0,32	0,68	0,32	0,58	17,86	15,11
	Faizabad	0,29	0,32	0,11	0,47	0,19	0,45	30,81	21,06
	Ambedkar Nr	0,31	0,34	0,18	0,44	0,22	0,46	20,62	17,70
	Sultanpur	0,21	0,24	0,11	0,31	0,15	0,32	24,39	19,03
	Bahraich	0,49	0,47	0,32	0,66	0,33	0,61	17,90	15,14
	Shrawasti	0,36	0,33	0,16	0,56	0,19	0,48	28,08	21,80
	Balrampur	0,20	0,23	0,05	0,35	0,11	0,35	39,42	26,68
	Gonda	0,27	0,29	0,15	0,39	0,18	0,40	22,81	18,76
	Siddharth Nr	0,26	0,26	0,14	0,38	0,15	0,36	23,43	20,87
	Basti	0,58	0,50	0,43	0,73	0,38	0,62	12,92	12,62
	Sant Kabir Nr	0,33	0,32	0,18	0,48	0,20	0,44	23,54	19,13
	Gorakhpur	0,28	0,27	0,17	0,39	0,18	0,37	19,90	17,88
	Kushinagar	0,21	0,23	0,10	0,32	0,13	0,32	26,53	21,49
	Deoria	0,35	0,33	0,20	0,50	0,21	0,45	21,49	18,38
	Azamgarh	0,32	0,32	0,21	0,43	0,21	0,42	18,15	16,69
	Mau	0,15	0,24	0,03	0,27	0,12	0,37	39,89	26,03
	Ballia	0,27	0,28	0,12	0,42	0,15	0,40	27,64	22,42
	Jaunpur	0,18	0,23	0,09	0,27	0,13	0,33	25,01	21,55
	Ghazipur	0,24	0,25	0,14	0,34	0,16	0,34	21,56	19,26
	Chandauli	0,19	0,20	0,06	0,32	0,10	0,30	34,87	25,62
	Varanasi	0,19	0,19	0,08	0,30	0,11	0,28	28,43	22,47
	S, R, Das Nr	0,51	0,41	0,34	0,68	0,28	0,54	17,40	16,46
	Mirzapur	0,24	0,24	0,14	0,34	0,14	0,34	22,06	21,25
	Sonbhadra	0,38	0,35	0,21	0,55	0,22	0,48	22,77	21,25

5 Observations en conclusion

Dans cet article, nous décrivons une extension spatiale de la version au niveau du domaine du MLGM et considérons l'estimation sur petits domaines des proportions pondérées selon l'enquête dans ce modèle. Nous introduisons la dépendance spatiale dans la structure des erreurs du MLGM par un processus autorégressif simultané (ARS). Nous examinons les différentes façons d'intégrer la dépendance spatiale résultante aux estimations sur petits domaines. Nous faisons intervenir une taille et un chiffre d'échantillon efficaces pour tenir compte du plan de sondage. Nous décrivons également un estimateur EQM analytique. Nous appliquons ensuite ces méthodes d'EPD à des données réelles d'enquête pour estimer la fréquence de la pauvreté en milieu rural et produire une carte correspondante de la pauvreté dans les différents districts

de l'État d'Uttar Pradesh en Inde. Nous évaluons nos résultats empiriques par plusieurs mesures de diagnostic et démontrons que les méthodes d'EPD par modèle font faire des gains d'efficacité dans l'estimation au niveau des districts des proportions de ménages pauvres. Ajoutons que, lorsque nous utilisons la corrélation spatiale entre les districts d'une manière appropriée, les estimations par modèle mènent à des gains significatifs d'efficacité des estimations au niveau des districts. Dans le cas de ces données, une mesure de proximité spatiale définie comme une fonction gaussienne des distances entre les districts semble le meilleur moyen de tenir compte de l'association spatiale entre ceux-ci. Disons enfin que l'exploitation de l'information d'enquête par l'emploi d'une taille d'échantillon efficace permet des estimations plus représentatives et réalistes.

Dans le contexte de l'estimation des dénombrements de petits domaines dans des modèles de petits domaines au niveau des unités, D'Alò, Di Consiglio, Falorsi, Ranalli et Solari (2012) décrivent plusieurs façons d'inclure dans l'analyse l'information spatiale par différentes mesures de distance. Il serait donc bon de sonder ces possibilités d'introduire l'information spatiale dans les versions au niveau du domaine des modèles correspondants de petits domaines. Dans notre article, nous employons également une structure d'autorégression simultanée pour appréhender la dépendance spatiale dans le MLGM. Comme nous l'avons indiqué, d'autres moyens comme les modèles autorégressifs conditionnels et les modèles nouvellement conçus de non-stationnarité spatiale peuvent servir à caractériser cette dépendance. Voir, par exemple, Besag et coll. (1991), Leroux et coll. (1999) et Chandra et coll. (2017, 2018). Ce sont des méthodes à explorer et nous sommes en train de les étudier.

Il convient enfin de noter que les estimations au niveau des districts de la fréquence de la pauvreté en milieu rural par les méthodes que nous avons décrites seront utiles à divers ministères du gouvernement indien et aux organisations internationales dans leur planification stratégique et leur recherche en politiques. Elles faciliteront aussi les affectations budgétaires et le ciblage des interventions en matière de bien-être en faisant connaître les districts ou les régions se caractérisant par une haute fréquence de la pauvreté rurale. Notre application est enfin la preuve que l'EPD peut constituer un moyen rentable et efficace de production de statistiques fiables désagrégées à partir des données d'enquête existantes grâce à une combinaison de l'information auxiliaire tirée de différentes sources publiées et des estimations directes d'enquête.

Remerciements

Les auteurs remercient de leurs observations et leurs suggestions utiles le rédacteur en chef, le rédacteur adjoint et les trois examinateurs. Elles auront permis d'améliorer considérablement notre article. Hukum Chandra a réalisé cette étude dans le cadre d'un projet national pour attachés de recherche de l'ICAR-IASRI à New Delhi en Inde. Le travail de Nicola Salvati a été subventionné par le *Progetto di Ricerca di Ateneo* du *Survey-based to register-based statistics: a paradigm shift using latent variable models* (subvention PRA2018-19).

Bibliographie

Anselin, L. (1992). *Spatial Econometrics*. Methods and Models. Kluwer Academic, Boston.

Banerjee, S., Carlin, B. et Gelfand, A. (2004). *Hierarchical Modelling and Analysis for Spatial Data*. New York: Chapman and Hall.

Besag, J., York, J. et Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43, 1-59.

Breslow, N.E., et Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistics Association*, 88(421), 9-25.

Brown, G., Chambers, R., Heady, P. et Heasman, D. (2001). Évaluation des méthodes d'estimation régionale dans leur application aux estimations du chômage tirées de l'Enquête sur la population active au Royaume-Uni. Recueil : *Symposium 2001, La qualité des données d'un organisme statistique : une perspective méthodologique*, Statistique Canada.

Chandra, H. (2013). Exploring spatial dependence in area level random effect model for disaggregate level crop yield estimation. *Journal of Applied Statistics*, 40, 823-842.

Chandra, H., et Salvati, N. (2018). Small area estimation for count data under a spatial dependent aggregated level random effects model. *Communications in Statistics - Theory and Methods*, 47(5), 1234-1255.

Chandra, H., Chambers, R. et Salvati, N. (2012). Small area estimation of proportions in business surveys. *Journal of Statistical Computation and Simulation*, 82(6), 783-795.

Chandra, H., Salvati, N. et Chambers, R. (2017). Small area prediction of counts under a non-stationary spatial model. *Spatial Statistics*, 20, 30-56.

Chandra, H., Salvati, N. et Chambers, R. (2018). Small area estimation under a spatially non-linear model. *Computational Statistics and Data Analysis*, 126, 19-38.

Chandra, H., Salvati, N. et Sud, U.C. (2011). Disaggregate-level estimates of indebtedness in the state of Uttar Pradesh in India-an application of small area estimation technique. *Journal of Applied Statistics*, 38(11), 2413-2432.

Cressie, N. (1993). *Statistics for Spatial Data*. New York: John Wiley & Sons, Inc.

D'Alò, M., Di Consiglio, L., Falorsi, S., Ranalli, M.G. et Solari, F. (2012). Use of spatial information in small area models for unemployment rate estimation at sub-provincial areas in Italy. *Journal of the Indian Society of Agricultural Statistics*, Special issue on small area estimation, 66(1), 43-54.

Fay, R.E., et Herriot, R. (1979). Estimates of income for small places: An application of James Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Fotheringham, A.S., Brunsdon, C. et Charlton, M.E. (2002). *Geographically Weighted Regression*. New York: John Wiley & Sons, Inc., West Sussex.

Franco, C., et Bell, W.R. (2013). Applying bivariate binomial/logit normal models to small area estimation. *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, 690-702.

Ghosh, M., Natarajan, K., Stroud, T.W.F. et Carlin, B.P. (1998). Generalized linear model for small-area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Jiang, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning Inference*, 111, 117-127.

Johnson, F.A., Chandra, H., Brown, J. et Padmadas, S. (2010). Estimating district-level births attended by skilled attendants in Ghana using demographic health survey and census data: An application of small area estimation technique. *Journal of Official Statistics*, 26(2), 341-359.

Korn, E.L., et Graubard, B.I. (1998). Intervalles de confiance pour les proportions à petit nombre d'événements positifs prévus estimées au moyen des données d'enquête. *Techniques d'enquête*, 24, 2, 209-218. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1998002/article/4356-fra.pdf.

Leroux, B., Lei, X. et Breslow, N. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. Dans *Statistical Models in Epidemiology, the Environment and Clinical Trials*, (Éds., M. Halloran et D. Berry), 135-178. New York: Springer-Verlag.

Liu, B., Lahiri, P. et Kalton, G. (2014). Modélisation hiérarchique bayésienne de proportions dans de petits domaines pondérées par les poids de sondage. *Techniques d'enquête*, 40, 1, 1-15. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014001/article/14030-fra.pdf.

Lopez-Vizcaino, E., Lombardìa, M. et Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statististcal Modelling*, 13, 153-178.

Malec, D., Sedransk, J., Moriarity, C.L. et LeClere, F.B. (1997). Small area inference for binary variables in the national health interview survey. *Journal of the American Statistical Association*, 92, 815-826.

Manteiga, G.W., Lombardia, M.J., Molina, I., Morales, D. et Santamarìa, L. (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.

Marhuenda, Y., Molina, I. et Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 58, 308-325.

McGilchrist, C.E. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society* B, 56, 61-69.

Mercer, L., Wakefield, J., Chen, C. et Lumley, T. (2014). A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spatial Statistics*, 8, 69-85.

Molina, I., Saei, A. et José Lombardìa, M. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *J. R. Statist. Soc. A*, 170, 975-1000.

Molina, I., Salvati, N. et Pratesi, M. (2009). Bootstrap for estimating the MSE of the spatial EBLUP. *Computational Statistics*, 24, 441-458.

Pfeffermann, D. (2002). Small area estimation: New developments and directions. *Revue Internationale de Statistique*, 70, 125-143.

Porter, A.T., Wikle, C.K. et Holan, S.H. (2015). Small area estimation via multivariate Fay-Herriot models with latent spatial dependence. *Australian and New Zealand Journal of Statistics*, 57, 15-29.

Pratesi, M., et Salvati, N. (2008). Small area estimation - the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, 17, 114-131.

Pratesi, M., et Salvati, N. (2009). Small area estimation in the presence of correlated random area effects. *Journal of Official Statistics*, 25, 37-53.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*, 2nd Edition. New York: John Wiley & Sons, Inc.

Saei, A., et Chambers, R. (2003). Small area estimation under linear and generalized linear mixed models with time and area effects. *Southampton Statistical Sciences Research Institute*, documents de travail de la méthodologie S3RI, M03/15.

Salvati, N., Chandra, H. et Chambers, R. (2012). Model based direct estimation of small area distributions. *Australian & New Zealand Journal of Statistics*, 54(1), 103-123.

Särndal, C.-E., Swensson, B. et Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Singh, B.B., Shukla, G.K. et Kundu, D. (2005). Modèles spatio-temporels pour l'estimation pour petits domaines. *Techniques d'enquête*, 31, 2, 201-214. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9053-fra.pdf.

Tzavidis, N., Salvati, N., Pratesi, M. et Chambers, R. (2008). M-quantile models with application to poverty mapping. *Statistical Methods & Applications*, 17(3), 393-411.