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

Techniques d’enquête

Indicateurs de qualité pour évaluer la représentation dans les données administratives

par Natalie Shlomo et Myong Sook Kim

Date de diffusion : le 23 décembre 2025

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 infostats@statcan.gc.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

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.

Techniques d'enquête, décembre 2025
Vol. 51, nº 2, pp. 555-581
Statistique Canada, nº 12-001-X au catalogue

Indicateurs de qualité pour évaluer la représentation dans les données administratives

Natalie Shlomo et Myong Sook Kim[1]

Résumé

Les organismes nationaux de statistique consacrent des ressources à la promotion de l'utilisation des données administratives dans les statistiques officielles. Cependant, les données administratives ne sont pas élaborées en vue de produire des statistiques. Il s'agit plutôt du résultat d'un événement ou d'une opération en lien avec des procédures administratives d'organisations, d'administrations publiques et d'organismes gouvernementaux. Il est donc essentiel d'évaluer la qualité des données administratives en ce qui concerne les sources d'erreur, tout particulièrement la représentativité par rapport à la population cible. Dans le présent document, on exploite la force des échantillons probabilistes de référence ou des recensements qui peuvent servir à déceler le manque de représentativité dans les données administratives, et l'on présente des indicateurs de qualité fondés sur des mesures de distance et des indicateurs de représentativité (indicateurs R). On fait état de leur application à l'aide d'une étude par simulation et l'on se penche sur une application réelle en utilisant un ensemble de données administratives du bureau national de la statistique du Royaume-Uni.

Mots-clés : Indicateurs R; mesures de la distance; scores de propension.

1. Introduction

Les organismes nationaux de statistique (ONS) consacrent des ressources à la promotion de l'utilisation des données administratives dans leurs systèmes statistiques nationaux. Il s'agit d'une priorité absolue pour le bureau national de la statistique (BNS) du Royaume-Uni, qui transforme ses systèmes statistiques afin d'utiliser davantage les données administratives pour les recensements à venir et la production de données sur la population. Par données administratives, on entend des sources de données secondaires, car ces données sont produites par d'autres organismes et sont le résultat d'un événement ou d'une opération en lien avec des procédures administratives d'organisations, d'administrations publiques et d'organismes gouvernementaux. Par ailleurs, les données administratives peuvent être améliorées en ayant davantage recours à la technologie de l'information et des communications, et elles sont de plus en plus facilement accessibles en format électronique. Elles pourraient donc devenir des sources de données de plus en plus importantes pour la production de statistiques officielles en réduisant considérablement le coût et le fardeau de réponse, et en améliorant l'efficacité des systèmes de production statistique.

L'ingestion de données administratives dans les systèmes statistiques des ONS n'est pas sans coût. Il est essentiel de comprendre où des erreurs potentielles peuvent se produire. Les ONS ont investi dans l'élaboration de cadres de qualité pour intégrer des sources de données externes, comme des données administratives, dans leurs systèmes statistiques. Bon nombre de ces cadres sont principalement de nature qualitative. L'un des premiers cadres de la qualité pour les données administratives dans le contexte européen a été élaboré par Daas, Ossen, Vis-Visschers et Arends-Toth (2009). Le cadre propose une liste

1. Natalie Shlomo, Social Statistics Department, School of Social Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, Royaume-Uni. Courriel : natalie.shlomo@manchester.ac.uk; Myong Sook Kim, Social Statistics Department, School of Social Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, Royaume-Uni. Courriel : myongsook.kim@postgrad.manchester.ac.uk.

de contrôle pour l'évaluation de la qualité des sources de données administratives en fonction d'« hyperdimensions » selon la source, les métadonnées et les données. Chaque hyperdimension peut avoir de nombreuses dimensions. Un indicateur de qualité peut être mesuré à l'aide de techniques qualitatives ou quantitatives. Par exemple, la dimension de la qualité des données manquantes pour une variable peut être exprimée comme la proportion de données manquantes. Depuis, de nombreux ONS ont élaboré des cadres de qualité semblables pour les données administratives. Parmi les exemples, citons l'assurance de la qualité des données administratives fondée sur les données du registre des patients du service national de la santé du Royaume-Uni. Le BNS a évalué ces données afin de produire des statistiques démographiques. Les auteurs proposent des mesures à suivre pour évaluer la qualité de ces données dans le cadre de vérifications du traitement et de la validation, en plus de suggérer des mesures pour améliorer l'ingestion de données (BNS, 2016). Plus récemment, un groupe d'étude de statisticiens européens chargé d'évaluer la qualité de sources administratives en vue de leur utilisation dans le cadre de recensements a publié un ensemble de lignes directrices qui ont été approuvées lors de la 69^{e} réunion plénière de la Conférence des statisticiens européens en 2021 (Commission économique des Nations Unies pour l'Europe, 2021). Les lignes directrices énoncent quatre hyperdimensions, soit l'étape de la source, l'étape des données, l'étape du traitement et l'étape de la production. Des dimensions relatives à la qualité figurent sous chaque hyperdimension et peuvent être évaluées de manière qualitative ou quantitative. Sur les traces de ce groupe d'étude, Dunstan et Correia (2021) ont élaboré un cadre de qualité sur mesure afin de garantir la qualité des données administratives utilisées dans le cadre du recensement de 2021 en Angleterre et au pays de Galles. Dans le présent document, le cadre de qualité compte deux hyperdimensions, soit l'étape de la source et l'étape des données. Ce cadre propose un avantage, à savoir que toutes les dimensions définies dans chacune des hyperdimensions ont été évaluées à l'aide d'une échelle de type Likert, par exemple 1) « ne correspond pas », 2) « correspond en partie » et 3) « correspond ». Les cotes peuvent donc être agrégées pour obtenir des cotes de qualité quantitatives globales pour chaque hyperdimension (Spector, 1992). De nombreux ONS consacrent des ressources à l'élaboration de cadres de qualité pour intégrer des données administratives en fonction de principes internationaux, comme ceux établis par les Nations Unies (2019). Ces cadres de qualité ont un point en commun, c'est-à-dire qu'ils comprennent les éléments fondamentaux des définitions des principes de qualité : la pertinence; l'exactitude et la fiabilité; l'actualité et la ponctualité; la cohérence et la comparabilité; et l'accessibilité (Eurostat, 2019).

Sur le plan conceptuel, le cycle de vie des microdonnées statistiques intégrées du point de vue de la qualité a été présenté dans Zhang (2012), où les sources d'erreur possibles sont relevées lors de l'utilisation de données administratives, en fonction du fait qu'il s'agit d'une source de données unique ou intégrée à d'autres sources de données, comme des données d'enquête. Dans le cas de données administratives d'une source unique, la représentation de la population cible d'intérêt fait partie des principales sources d'erreur. La présente étude porte principalement sur cette dimension de la qualité.

On suggère des indicateurs de la qualité qui permettent aux ONS d'évaluer, de manière quantitative, si les données administratives sont représentatives de la population cible et de déterminer les sous-groupes qui

peuvent ne pas être visés par les données administratives. En outre, les données administratives sont souvent fournies sous forme de flux. Les ONS doivent savoir à quel moment les données administratives sont suffisamment exhaustives. Donc, il est nécessaire d'évaluer constamment les données administratives, afin de déterminer à quel moment les données sont exhaustives et peuvent être utilisées dans le cadre de la production.

Pour évaluer la qualité des données administratives, les ONS devraient tirer parti des données récentes des recensements ou des enquêtes probabilistes (pondérées) de grande qualité pour comparer les distributions entre cette source de données auxiliaires et les données administratives. Puisque les données administratives peuvent être ingérées plus fréquemment par rapport aux données d'un recensement tenu tous les 5 ou 10 ans, les ONS devraient utiliser les collectes de données d'enquête pour évaluer la qualité des données administratives. Par exemple, de vastes collectes de données, comme le Labour Force Survey, peuvent servir à évaluer la qualité des données administratives en ce qui concerne la représentativité. Même si l'on sait que ces enquêtes peuvent être associées à un taux élevé de non-réponse, elles sont habituellement de bonne qualité en ce qui concerne les mesures prises pour atténuer le biais de non-réponse pendant la collecte des données et les corrections postérieures à l'enquête qui compensent la non-réponse. En outre, les ONS réalisent habituellement une poststratification et assurent le calibrage des facteurs de pondération du plan d'enquête corrigés pour tenir compte de la non-réponse par rapport à des repères démographiques connus, qui sont généralement le sexe, l'âge et la région géographique. Ces repères démographiques sont habituellement obtenus chaque année par les ONS à l'aide des principes de comptabilité démographique entre les années de recensement et sont traités comme de « vraies » valeurs.

Parmi les autres exemples de données d'enquête de grande qualité qui peuvent servir à évaluer la représentativité des données administratives, il y a les données d'enquête recueillies dans le but unique de corriger la couverture. Dans de nombreux pays, les enquêtes sur la couverture sont obligatoires (semblables à un recensement). Elles doivent donc être de grande qualité et peuvent servir à évaluer la qualité des données administratives. Il convient de souligner que les enquêtes sur la couverture sont généralement utilisées pour l'estimation de système dual afin de tenir compte du sous-dénombrement lors d'un recensement, par exemple en permettant l'estimation des poids de couverture pour le recensement. C'est la raison pour laquelle ces types d'enquêtes sur la couverture permettent d'avoir des estimations judicieuses des distributions dans la population quand vient le temps d'évaluer la qualité des données administratives. Dans le présent document, on suppose que l'on a accès à un échantillon probabiliste de référence de grande qualité pour évaluer la représentativité des données administratives.

On utilise des distributions univariées et bivariées obtenues à partir d'un recensement ou d'une enquête probabiliste pondérée de référence, et on les compare à des distributions dans les données administratives en fonction d'un ensemble commun de variables. On peut utiliser des mesures de distance pour évaluer la différence entre les distributions. En outre, on suggère un indicateur de la représentativité (R) conçu pour quantifier la représentativité des groupes de population dans les données administratives par l'estimation d'un score de propension et l'étude de sa variation. On considère le score de propension comme un score

d'équilibrage selon lequel, sous réserve du score de propension, la distribution des covariables de base observées devrait être similaire entre les données administratives et les données pondérées d'enquête ou de recensement (Austin, 2011).

L'approche suggérée repose sur de l'information auxiliaire agrégée sous forme de tableaux statistiques au niveau des variables et des catégories pour la base de l'estimation des scores de propension fondés sur une fonction de régression linéaire. Il convient de souligner que cette approche est semblable à l'approche par calage des poids et à d'autres indicateurs linéaires de qualité d'enquête élaborés, par exemple, dans Särndal et Lundström (2008). Cette approche évite d'avoir besoin de données individuelles au niveau des enregistrements pour estimer les scores de propension, comme on peut le voir, par exemple, dans Chen, Li et Wu (2020). En outre, comme cela arrive souvent lors de l'estimation des probabilités de couverture pour corriger le sous-dénombrement et le surdénombrement lors d'un recensement (Alho, 1990; BNS, 2022), le couplage au niveau des enregistrements entre les données administratives et la source de données auxiliaires n'est pas nécessaire, car on ne cherche pas à estimer une probabilité de couverture individuelle. Dans ce cas-ci, on cherche à fournir des indicateurs de qualité pour déterminer la représentativité au niveau d'un groupe plutôt qu'au niveau individuel. On dispose donc d'un avantage, c'est-à-dire que l'application est plus rapide et plus simple au niveau computationnel et qu'elle empêche l'introduction d'éventuelles erreurs de couplage. Cette approche a un désavantage : il faut utiliser un modèle de régression linéaire pour estimer les scores de propension plutôt qu'un modèle de régression logistique ou probit, ce qui fait en sorte que l'on pourrait obtenir des scores de propension estimés pouvant être hors de l'intervalle [0, 1]. Cependant, puisque l'on cherche principalement à évaluer la variation des scores de propension pour obtenir des indicateurs de qualité, on met en application cette approche pratique et simple à mettre en œuvre.

On montre la manière dont ces mesures de la qualité quantitatives peuvent servir à évaluer la représentativité dans les données administratives et peuvent fournir une évaluation plus directe de la qualité par rapport aux listes de contrôle et aux fiches de scores utilisées auparavant qui sont principalement de nature qualitative. De plus, l'estimation du score de propension dans les données administratives permet de corriger les estimations obtenues à partir des données administratives en les pondérant selon l'inverse du score de propension, de manière semblable à la pondération par l'inverse de la propension (PIP) pour les échantillons d'enquête mentionnée dans Kim et Haziza (2014) et, plus récemment, dans Chen et coll. (2020) et dans les références à cet égard dans le cas des enquêtes non probabilistes. Pour utiliser l'approche de PIP en présence d'échantillons non probabilistes, Chen et coll. (2020) énumèrent les hypothèses nécessaires pour produire des estimations sans biais : 1) l'indicateur d'inclusion des données administratives est indépendant de toute variable cible compte tenu d'un ensemble de covariables $\mathbf{X}$; 2) toutes les unités ont un score de propension non nul à ajouter aux données administratives; 3) les indicateurs d'inclusion pour les unités i et j sont indépendants compte tenu de x_i et de x_j pour $i \neq j$. Dans le cas des données administratives, les ONS vont harmoniser et agencer de multiples sources de données administratives dans le but d'assurer la couverture d'une population cible (voir l'application à la section 4). Les personnes qui ne sont pas présentes dans les données administratives ou qui ont des adresses erronées ou multiples sont habituellement les personnes qui

ne font pas d'opérations avec le gouvernement. Par exemple, les jeunes et les étudiants peuvent figurer à d'autres adresses. Malgré tout, on suppose que l'on peut expliquer tout écart systématique de la population cible par les covariables $\mathbf{X}$ dans les données et que les hypothèses de Chen et coll. (2020) dans le cas d'un échantillon non probabiliste seront valables également dans le contexte des données administratives.

À la section 2, on décrit les indicateurs de qualité suggérés pour évaluer la représentativité des données administratives. À la section 3, on montre les indicateurs de qualité suggérés pour les données administratives (simulées) et l'on donne une interprétation des résultats. À la section 4, on examine une application reposant sur un ensemble de données administratives réel mis en œuvre dans des serveurs sécurisés du BNS. On conclut le tout par une discussion à la section 5.

2. Indicateurs de qualité

Dans de nombreux cadres d'évaluation de la qualité des données administratives des ONS, on utilise très peu les données de recensement auxiliaires externes ou les données d'enquête de grande qualité recueillies par l'organisme, afin de quantifier la représentativité dans les données administratives.

Habituellement, les données administratives comportent les erreurs suivantes : des personnes sont associées à la mauvaise adresse; des enregistrements sont manquants parce que les personnes peuvent ne pas interagir avec l'organisme; des enregistrements sont en double, car des personnes peuvent être associées à différentes adresses. En outre, il peut y avoir des enregistrements erronés, comme les décès et les émigrants qui ont quitté le pays. Les ONS utilisent des pratiques exemplaires pour éliminer les doublons et nettoyer les données administratives avant qu'elles soient utilisées dans les systèmes statistiques. Ils peuvent supprimer les décès et les émigrants enregistrés des données administratives, bien qu'il soit difficile de cerner et d'éliminer les émigrants non enregistrés. Dans la présente étude, on suppose que le nombre d'enregistrements dans les données administratives correspond à M, que le nombre d'enregistrements (pondérés) dans les données de référence correspond à N et qu'il est possible que M soit supérieur à N.

Dans la présente section, on montre comment il est possible d'utiliser ce type de données auxiliaires pour créer des indicateurs de la qualité des données administratives. On trouve le logiciel de code R (R Core Team, 2022) pour les indicateurs de qualité et un guide de l'utilisateur sur GitHub (Kim et Shlomo, 2023).

2.1 Mesures de la distance

Dans la présente section, on compare les distributions univariées et bivariées obtenues à partir des données administratives aux données auxiliaires agrégées de population externes, qu'elles aient été obtenues directement dans le cadre d'un recensement ou qu'elles aient été estimées à partir d'une enquête probabiliste aléatoire de grande qualité.

Désignons la variable k ayant des catégories $h, h = 1, 2 \ldots H$ dans l'ensemble de données administratives comprenant M personnes. Supposons que $\Delta_{h,i}^{k}$ est l'indicateur $0-1$ pour la personne i qui est membre de la catégorie h de la variable k. On peut ensuite calculer les chiffres pour la catégorie h : $m_h^k = \sum_{i=1}^{M} \Delta_{h,i}^{k}$ et

prendre note que $\sum_{h=1}^{H} m_h^k = M$. On peut également obtenir la distribution probabiliste de la variable k ayant la catégorie $h, h = 1,2 \ldots H$ et calculer $p_h^k = \frac{m_h^k}{M}$. La variable k peut également représenter un tableau croisé d'au moins deux variables, par exemple le groupe d'âge $\times$ le sexe.

Supposons maintenant que l'on a obtenu des estimations équivalentes de ces distributions d'un recensement ou, sinon, d'un grand échantillon probabiliste aléatoire de taille n dans lequel toutes les personnes i de l'échantillon se sont vu attribuer un poids de sondage connexe w_i. Habituellement, les poids de sondage sont corrigés pour tenir compte de la non-réponse et calibrés en fonction de repères démographiques connus, de sorte que leur somme corresponde à la taille connue de la population N. Dans ce cas-ci, $n_h^k = \sum_{i=1}^{n} w_i \Delta_{h,i}^k$, $\sum_{h=1}^{H} n_h^k = N$ et la distribution probabiliste équivalente est $q_h^k = \frac{n_h^k}{N}$.

On peut maintenant définir différentes mesures de la distance pour évaluer les écarts entre les distributions $\{p_h^k, h = 1, \ldots, H\}$ et $\{q_h^k, h = 1, \ldots, H\}$. Dans la présente étude, on montre deux mesures de la distance pour la variable k (on supprime l'indicateur pour la variable k), soit l'indicateur de dissimilarité (ID) (Duncan et Duncan, 1955) :

$$\text{ID} = \frac{1}{2} \sum_h \left| p_h - q_h \right|,$$

et la distance de Hellinger (DH) :

$$\text{DH} = \frac{1}{\sqrt{2}} \sqrt{\sum_h \left(\sqrt{p_h} - \sqrt{q_h} \right)^2}.$$

Il existe de subtiles différences entre les mesures de distance. Par exemple, la distance de Hellinger accorde plus de poids aux proportions plus petites par rapport aux proportions plus grandes, alors que l'indicateur de dissimilarité traite de manière égale toutes les proportions. On a également évalué la divergence de Kullback-Leibler. Cependant, elle est associée à la distance de Hellinger et n'est pas une « vraie » mesure de la distance (elle ne respecte pas l'inégalité triangulaire). On a ajouté le calcul des mesures de distance dans le code R sur GitHub (Kim et Shlomo, 2023). Ce calcul facilitera la réalisation de travaux plus empiriques pour les recommandations futures.

Les deux mesures de distance, soit la DH et l'ID, sont bornées par 0 pour une concordance parfaite dans la distribution et par 1 pour une non-concordance parfaite. Une non-concordance parfaite se produit lorsque la totalité de la masse des deux distributions se trouve dans des cellules différentes de la distribution. Dans le cadre de la présente étude, on définit 1 moins la mesure de distance, que l'on désigne par $(1 - \text{DH})$ et $(1 - \text{ID})$, afin que 1 représente une concordance parfaite et que 0 représente une non-concordance parfaite.

Pour de petites mesures $(1 - \text{DH})$ ou $(1 - \text{ID})$, on peut déterminer la catégorie de la variable à laquelle est attribuable la petite mesure de distance et tenter de corriger le manque de représentativité en désignant d'autres sources de données qui s'ajouteront aux données administratives. Il convient de souligner que ces mesures de la distance peuvent quantifier des différences au chapitre des statistiques univariées et bivariées. Cependant, pour obtenir une évaluation multivariée plus détaillée des variables, on se tourne maintenant vers l'utilisation des indicateurs R à la section 2.2.

2.2 Indicateurs R

Au départ, l'indicateur R et ses indicateurs R partiels connexes ont été conçus pour évaluer la représentativité des réponses d'une enquête. Ils sont particulièrement utiles comme fonction objective lors de l'optimisation des plans de sondage adaptatifs (Schouten, Cobben et Bethlehem, 2009; Schouten et Shlomo, 2017). Les indicateurs R mesurent le contraste entre les personnes qui sont absentes ou non dans les données et permettent de déterminer les groupes qui ne sont pas représentés dans les données recueillies. Dans le cadre d'enquêtes, on évalue les propensions à répondre à l'aide d'un modèle de régression logistique dans les microdonnées d'enquête en fonction des covariables accessibles à partir de la base de sondage, des données auxiliaires et des paradonnées. Dans le cadre du modèle de régression logistique (distribution binomiale), l'écart-type maximum des propensions à répondre estimées est de 0,5. Donc, dans la documentation fondée sur les enquêtes, l'indicateur R est défini comme étant $\hat{R}_{\hat{\rho}} = 1 - 2 \times \mathrm{ET}_{\hat{\rho}}$, où $\mathrm{ET}_{\hat{\rho}}$ est l'écart-type des propensions à répondre estimées. Cela permet d'avoir un indicateur R borné entre (près de) 0 et 1. Voir Schouten et coll. (2009) pour en savoir davantage.

Ouwehand et Schouten (2014) ont élaboré des indicateurs R pour évaluer les statistiques conjoncturelles tirées des enquêtes auprès des entreprises ainsi que leur représentativité dans la source des données administratives du registre de la taxe sur la valeur ajoutée pour les entreprises aux Pays-Bas. Dans ce cas-ci, on a utilisé les données administratives comme source de données auxiliaires. Les auteurs ont utilisé l'approche standard visant à évaluer les indicateurs R en estimant des scores de propension selon la documentation portant sur l'échantillon mentionnée ci-dessus.

Dans la présente étude, on élabore l'indicateur R et les indicateurs R partiels, afin d'évaluer la représentativité d'un ensemble de données administratives par rapport à une population cible et l'on utilise une approche différente. Dans une étude récente de Bianchi, Shlomo, Schouten, Da Silva et Skinner (2019), l'indicateur R est adapté en fonction du cas dans lequel on souhaite évaluer la représentativité des données d'enquête, mais que seule l'information auxiliaire agrégée de population est accessible, au lieu de renseignements de la base échantillonnale. Cette approche a été mise en application avec succès dans Shlomo, Luiten et Schouten (2022), où les auteurs ont évalué la représentativité des statistiques sur le revenu et les conditions de vie dans de multiples pays de l'Union européenne. On s'inspire de cette étude et l'on utilise l'information auxiliaire agrégée de population obtenue lors d'un récent recensement ou de dénombrements d'échantillon pondérés calculés à partir d'un grand échantillon probabiliste aléatoire de grande qualité.

Pour calculer l'indicateur R basé sur la population, désignons l'indicateur d'inclusion r_i comme égal à un pour toutes les unités figurant dans l'ensemble de données administratives (désigné par « A »). On dispose de renseignements sur les valeurs $\mathbf{x}_i = (x_{1,i}, x_{2,i}, \ldots, x_{K,i})^T$ d'un vecteur de K variables auxiliaires $\mathbf{X}$, par exemple le sexe, le groupe d'âge, la région géographique, le groupe de minorité ethnique et la situation vis-à-vis de l'activité économique. Ainsi, chaque $x_{k,i}$ est une variable indicatrice binaire. On suppose également que les valeurs de $\mathbf{x}_i$ sont observées chez toutes les personnes de l'ensemble de données administratives, de sorte que l'on observe $\{x_i; i \in A\}$.

Supposons que l'on connaît $\mathbf{x}_i$ au niveau agrégé en fonction d'un ensemble de données externes, que ce soit à partir d'un recensement ou d'un échantillon probabiliste aléatoire. Pour calculer les estimations des scores de propension, on a besoin du total de population $\sum_U \mathbf{x}_i$ et des produits croisés de population $\sum_U \mathbf{x}_i \mathbf{x}_i^T$, où U représente la population. Il s'agit de l'information auxiliaire de population. En supposant que l'on a accès à un échantillon probabiliste aléatoire, soit s, où chaque personne i dans l'échantillon a un poids d'enquête w_i qui est corrigé et calibré en fonction de la taille de la population N, on peut envisager deux types de calculs pour estimer les produits croisés de population $\sum_U \mathbf{x}_i \mathbf{x}_i^T$ (Bianchi et coll., 2019) :

Type 1 : Si l'échantillon probabiliste aléatoire est vaste ou si l'on a accès à des données de recensement récentes, on peut estimer l'information auxiliaire de population à l'aide de : $\sum_U \mathbf{x}_i \approx \sum_s w_i \mathbf{x}_i$ et $\sum_U \mathbf{x}_i \mathbf{x}_i^T \approx \sum_s w_i \mathbf{x}_i \mathbf{x}_i^T$.

Type 2 : Si l'échantillon probabiliste aléatoire est petit et s'il est impossible d'estimer les produits croisés à partir de l'échantillon, on peut compter sur l'information marginale comme suit :

Étape 1 : Calculer une estimation pour la moyenne de la population à partir de l'échantillon : $\bar{X}_U = \frac{\sum_U \mathbf{x}_i}{N} \approx \frac{\sum_s w_i x_i}{N}$.

Étape 2 : Estimer les produits croisés dans la population à l'aide de $N\hat{S}_{XX} + N\bar{X}_U \bar{X}_U^T$, où l'on estime que $\hat{S}_{XX} = \left(\sum_{i=1}^M d_i\right)^{-1} \sum_{i=1}^M d_i (\mathbf{x}_i - \bar{X}_U)(\mathbf{x}_i - \bar{X}_U)^T$ à partir des données administratives A de taille M et où d_i est un poids d'ajustement pour les données administratives, appelé poids de représentativité ci-dessous, de sorte que $\sum_{i=1}^M d_i = N$.

On définit le score de propension pour une unité dans l'ensemble de données administratives comme l'espérance conditionnelle de la variable indicatrice r_i compte tenu des valeurs des covariables précisées. On désigne $\rho_X(\mathbf{x}_i)$ par ρ_i.

Dans le cadre fondé sur la population, on modélise les scores de propension selon une fonction de lien identité (linéaire) où les vrais scores de propension satisfont à $\rho_i = \mathbf{x}_i^T \boldsymbol{\beta}$, $i \in A$. Si l'on suppose que les données administratives ont un mécanisme de sélection probabiliste (bien qu'il s'agisse de probabilités inconnues), on peut écrire l'espérance conditionnelle de l'indicateur r_i compte tenu de la $\mathbf{x}_i$ auxiliaire de la façon suivante : $E(r_i \mid \mathbf{x}_i) = \rho_X(\mathbf{x}_i) \equiv \rho_i$. L'utilisation de la fonction de lien linéaire permet l'utilisation d'estimations de substitution directement dans la formule servant à estimer les scores de propension. Dans le cas du modèle probabiliste linéaire, l'estimation de ρ_i dans les données administratives A est fournie par : $\hat{\rho}_i^{\text{MCO}} = \mathbf{x}_i^T \hat{\boldsymbol{\beta}} = \mathbf{x}_i^T \left(\sum_A d_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_A d_i \mathbf{x}_i$, $i \in A$, où d_i est le poids de représentativité de l'unité i.

On utilise les estimateurs par substitution pour approximer $\hat{\rho}_i^{\text{MCO}}$ par $\hat{\rho}_i^P = \mathbf{x}_i^{\text{T}} \left(\sum_U \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_A d_i \mathbf{x}_i$, $i \in A$. Il faut souligner que l'on calcule $\hat{\rho}_i^P$ uniquement pour l'ensemble de personnes figurant dans les données administratives A.

L'estimation de $\hat{\rho}_i^{\text{MCO}}$ repose sur une approche semblable à la méthode fondée sur la calibration pour estimer les scores de propension d'un échantillon non probabiliste lorsque seuls les totaux de population sont connus en résolvant les équations d'estimation $\sum_{\text{NP}} \frac{x_i}{\pi(x_i,\theta)} - \sum_U x_i = 0,$ comme le montrent Chen, Li, Rao et Wu (2022). Les auteurs soulignent que les estimateurs basés sur la calibration et les estimateurs du maximum de pseudo-vraisemblance pour les paramètres du modèle, comme le montrent Chen et coll. (2020), ne sont pas équivalents sur le plan mathématique. Cependant, les deux estimateurs sont cohérents si l'on tient compte de la forme paramétrique pour estimer les scores de propension.

Dans notre cadre pour évaluer la représentativité dans les données administratives, on utilisera un seul poids représentatif : $d_i \equiv d = [M/N]^{-1}$ et est égal pour toutes les personnes $i \in A$ (n'oublions pas que M est la taille des données administratives et que N est la taille de la population). En appliquant d à tous les enregistrements dans les données administratives, on veille à obtenir des totaux globaux égaux. On ne calcule pas des poids représentatifs plus détaillés au niveau des strates, car on souhaite déterminer les variables et leurs catégories qui favorisent le manque de représentativité.

Pour obtenir l'indicateur R, il faut estimer la variance des scores de propension au niveau de la population : $S_\rho^2 = \frac{1}{N-1}\sum_U (\rho_i - \overline{\rho})^2$. Dans le cadre fondé sur la population, en fonction des données administratives accessibles, on peut estimer cette variance des scores de propension comme suit : $\hat{S}^2_{\hat{\rho}^P} = \frac{1}{N-1}\sum_A d_i (\hat{\rho}_i^P)^{-1}(\hat{\rho}_i^P - \hat{\overline{\rho}})^2$, où l'on estime $\hat{\rho}_i^P$ comme ci-dessus et où $\hat{\overline{\rho}}$ est le score de propension moyen global. Il convient de souligner que $(\hat{\rho}_i^P)^{-1}$ tient compte de tout biais associé au manque de représentation dans les données administratives. Comme cela est le cas dans Bianchi et coll. (2019), on estime la variance des scores de propension estimés à l'aide de : $\hat{S}^2_{\hat{\rho}^P} = \frac{N}{N-1}\left\{\frac{1}{N}\sum_A d_i \hat{\rho}_i^P - \left(\frac{1}{N}\sum_A d_i\right)^2\right\}$. Ce format permet la mise en œuvre de la correction du biais lié à la taille de l'échantillon. Cependant, cette correction n'est pas nécessaire pour évaluer la représentativité de vastes ensembles de données administratives.

S'il n'y a pas de variation au chapitre des scores de propension, la variance est (près de) 0, et les données administratives sont représentatives de la population cible. Plus la variation est grande, moins les données administratives de l'ensemble de variables auxiliaires sont représentatives.

L'utilisation de la fonction de lien identité permet d'obtenir des scores de propension estimés qui peuvent être supérieurs à une valeur de 1. Cela s'explique notamment par le fait que le poids de représentativité d est près de 1 et que les scores de propension sont susceptibles d'être élevés. Cependant, puisque l'on s'intéresse uniquement aux valeurs prédites pour évaluer la variation des scores de propension plutôt qu'à l'inférence et à l'interprétation des rapports de cotes et de l'ampleur des effets, on continue d'utiliser la fonction de lien identité pour élaborer des indicateurs de qualité basés sur les données administratives et l'on accepte que certains scores de propension soient supérieurs à 1.

Dans la documentation sur les enquêtes et dans le cadre d'une fonction de régression logistique, l'indicateur R estimé est défini comme $\hat{R}_{\hat{\rho}^P} = 1 - 2\hat{S}_{\hat{\rho}^P}$ (Schouten et coll., 2009), car l'écart-type maximum des scores de propension selon la régression logistique est de 0,5. Cependant, dans le cadre des données administratives, on définit l'estimateur de l'indicateur R comme suit : $\hat{R}_{\hat{\rho}^P} = 1 - \hat{S}_{\hat{\rho}^P}$. Cela s'explique par le fait que l'on fonde nos prévisions sur la distribution normale pour estimer les scores de propension pour les

données administratives. On s'attend à ce que le score de propension moyen soit élevé, ce qui suppose des écarts-types plus faibles par rapport à l'utilisation de la régression logistique modélisée selon la distribution binomiale. Si les scores de propension sont entièrement aléatoires, on peut s'attendre à un écart-type maximum de 1 selon la distribution normale et à un indicateur R (près de) 0. Si les scores de propension sont tous égaux, l'écart-type selon la distribution normale est de 0 et l'indicateur R est (près de) 1. On souligne que l'indicateur R est une fonction des variables auxiliaires dans le modèle de régression. Si l'on avait un ensemble différent de variables auxiliaires, on obtiendrait un indicateur R différent. En outre, le but de l'indicateur R est d'évaluer la représentativité sous forme d'« écarts » par rapport à une population cible. Il ne s'agit pas d'un but lié à l'inférence. Actuellement, on ne met pas en œuvre de diagnostics sur l'ajustement du modèle de score de propension. Cependant, ce sujet pourrait faire l'objet de travaux futurs.

Les indicateurs R partiels non conditionnels sont également mentionnés dans Bianchi et coll. (2019). Ces indicateurs mesurent l'ampleur de la variation des scores de propension entre les catégories d'une variable. Plus la variation entre les catégories est grande, plus l'incidence de la variable sur le manque de représentativité est élevée. Comme on l'a fait plus haut, désignons x_k comme l'une des composantes du vecteur $\mathbf{X}$. La variable x_k est catégorique avec H catégories (on met de côté l'indice k). Supposons que m_h désigne la taille pondérée dans les données administratives de la catégorie h pour $h=1, 2, \ldots, H$. Cela signifie que $m_h = \sum_{i \in A} d\Delta_{h,i}$, où $\Delta_{h,i}$ est l'indicateur $0-1$ pour l'unité participante i membre de la catégorie h et $\sum_{h=1}^{H} m_h = N$ en raison de la définition de d comme poids de la représentativité. Définissons $\hat{\bar{\rho}}_h$ comme score de propension moyen dans la catégorie h de x_k pour les unités dans l'ensemble de données administratives et $\hat{\bar{\rho}}$ comme score de propension moyen global en fonction des scores de propension estimés basés sur la population $\hat{\rho}_i^P$. L'estimation pour l'indicateur R partiel non conditionnel pour la variable x_k est : $R_U(x_k) = \sqrt{\frac{1}{N}\sum_{h=1}^{H} m_h(\hat{\bar{\rho}}_h - \hat{\bar{\rho}})^2}$. Plus la valeur de l'indicateur R partiel est élevée, plus l'association entre la variable et le manque de représentativité dans l'ensemble de données administratives est forte. En calculant et en comparant les indicateurs partiels non conditionnels d'un ensemble de variables, on peut établir pour quelles variables les relations avec le manque de représentativité sont les plus fortes. L'indicateur R partiel non conditionnel au niveau de la catégorie h pour la variable x_k est $R_U(x_k^h) = \sqrt{\frac{m_h}{N}}\,(\hat{\bar{\rho}}_h - \hat{\bar{\rho}})$ et peut être associé à des valeurs positives ou négatives. Il convient de souligner qu'au niveau de la catégorie, un signe négatif représente une sous-représentation et un signe positif représente une surreprésentation.

Enfin, on souligne que, lors de la production d'estimations à partir de l'ensemble de données administratives, il faut pondérer chaque personne i par l'inverse de son score de propension $(\hat{\rho}_i^P)^{-1}$ pour tenir compte du biais découlant du manque de représentation dans les estimations. Les distributions pondérées découlant des données administratives se rapprocheront des distributions de la population repères, surtout dans le cas des variables qui figurent dans le modèle pour estimer les scores de propension. On le verra à la section 3.

2.3 Estimation de la variance des indicateurs R

Il existe deux sources de variation pour les indicateurs de qualité suggérés et les estimations de la PIP calculées à partir des données administratives : 1) en raison de la grande taille des ensembles de données administratives représentant une population cible, il y aura habituellement de petites variances en ce qui concerne les estimations obtenues à partir des données administratives, même s'il y aura des sources supplémentaires de variation si les données administratives ont fait l'objet d'un processus de traitement des données, comme des ajustements à la couverture et une imputation; 2) si les repères démographiques utilisés pour estimer les scores de propension sont fondés sur des estimations pondérées calculées à partir de données d'enquête aléatoire de référence de grande qualité, on peut s'attendre à une variation supplémentaire attribuable à l'estimation de ces données repères.

Lorsque des chiffres univariés et bivariés sont accessibles à partir d'un recensement ou de totaux démographiques connus, et qu'ils servent de substitutions quand vient le temps d'estimer les scores de propension, il faut mettre l'accent sur la source de variation au point 1 ci-dessus. Dans ce cas-ci, pour estimer la variance des indicateurs R, des mesures de distance et des estimations de la PIP obtenues à partir des données administratives, on peut utiliser une approche de rééchantillonnage standard, comme l'approche bootstrap :

- Procéder à un échantillonnage avec remise à partir des données administratives;
- Mettre en œuvre toute procédure de traitement des données réalisée sur les données originales;
- Estimer les scores de propension pour B répétitions en fonction des totaux démographiques connus;
- Calculer l'estimateur d'intérêt et obtenir l'estimation de la variance en multipliant la somme des carrés pour les répétitions par 1/B.

On peut également ajouter aux données administratives un ensemble de poids de réplication bootstrap pour faciliter l'estimation de la variance pour les utilisateurs des données.

Comme on le mentionne au point 2 ci-dessus, l'incertitude liée aux estimations des indicateurs et aux estimations de la PIP peut augmenter lorsque l'on utilise des chiffres pondérés d'un échantillon de référence comme données repères pour estimer les scores de propension. Dans ce cas, la variance est $O(1/n)$ plutôt que $O(1/N)$, où n est la taille de l'échantillon de l'enquête. Wu (2022) fournit des expressions analytiques de la variance pour les estimations de la PIP associées à l'échantillonnage non probabiliste, ce qui peut être pertinent dans ce cas-ci. En outre, Opsomer et Erciulescu (2021) présentent une approche pour obtenir des poids de réplication découlant de l'approche bootstrap pour estimer la variance après une calibration en fonction de l'échantillon qui tient compte de la variation supplémentaire attribuable aux données repères obtenues à partir des échantillons d'enquête.

Par ailleurs, si les microdonnées de l'échantillon d'enquête, y compris les variables pertinentes, les variables du plan de sondage et les poids de sondage, sont accessibles, on peut utiliser une approche bootstrap à deux étapes. L'étude de Chen et coll. (2022) montre que les procédures bootstrap standard avec

remise appliquées séparément à l'échantillon et aux données administratives peuvent fournir une bonne couverture pour les intervalles de confiance lorsque l'échantillon probabiliste de référence est sélectionné à l'aide de plans d'échantillonnage selon des probabilités égales ou inégales à un seul degré.

Dans le contexte où l'on évalue la qualité de vastes ensembles de données administratives à l'aide d'un échantillon de référence pour les repères démographiques, il est possible que l'on n'ait pas accès aux microdonnées des enquêtes quand vient le temps de mettre en application l'approche bootstrap à deux étapes. Dans cette situation, le fait d'utiliser des enquêtes probabilistes de très grande envergure pour obtenir les repères démographiques, comme la Labour Force Survey, l'American Community Survey, les Coverage Surveys et autres, peut avoir une faible incidence sur une variation supplémentaire en raison de la grande taille des échantillons. En outre, le calcul des poids de sondage pour ces données d'enquête de grande qualité comprend habituellement l'étape de la poststratification, où les poids de sondage sont calibrés en fonction de totaux démographiques connus, comme cela est mentionné à la section 1, et n'ont donc pas de variance. Ainsi, la procédure bootstrap standard décrite ci-dessus, si l'on suppose que les données repères sont de « vrais » chiffres de population, pourrait être appliquée au cas dont il est question.

Dans l'étude par simulation, on montre les erreurs-types de simulation découlant du tirage de multiples échantillons aléatoires pour produire les repères démographiques. On calcule les scores de propension de l'ensemble de données administratives pour chaque échantillon, avant de montrer les erreurs-types de simulation dans les tableaux et les intervalles de confiance dans les figures à la section 3. Comme prévu, les erreurs-types de simulation sont petites. On montre également certains résultats associés à une approche d'échantillonnage bootstrap à deux étapes, comme celle utilisée dans Chen et coll. (2022), afin d'illustrer leur similitude par rapport aux erreurs-types de simulation dans le tableau 3.5 de la section 3.

3. Démonstration des indicateurs de qualité et simulation

Pour faire la démonstration des indicateurs de qualité suggérés, on utilise un ensemble simulé de microdonnées de recensement comptant six autorités locales anonymisées de trois régions du Royaume-Uni et produit à partir du recensement de 2021 au Royaume-Uni en réunissant de nombreuses données tabulaires publiées de celui-ci. Cela a entraîné la création d'un ensemble de données (synthétiques) de 1 163 659 personnes. Voici les variables simulées à partir de l'ensemble de données : l'élément géographique (autorité locale) (6 catégories), le sexe (2 catégories), le groupe d'âge (14 catégories), le groupe ethnique (16 catégories), l'état matrimonial (6 catégories) et le statut économique (10 catégories).

À partir de cet ensemble de données, on réalise deux étapes : 1) produire un ensemble de données administratives traité comme un ensemble fixe; 2) tirer B = 40 échantillons aléatoires sans remise à partir du recensement afin de les utiliser comme échantillons probabilistes de référence pour estimer les repères démographiques pour la modélisation du score de propension. Dans le cas des échantillons aléatoires, on tire un échantillon de 1:50 et l'on ne suppose aucune non-réponse, c'est-à-dire que chaque personne figurant dans l'échantillon a un poids de sondage de 50. On répète la simulation B = 40 fois avec chaque échantillon

aléatoire et l'on montre, dans la section des résultats, les résultats moyens et les erreurs-types de simulation (intervalles de confiance) obtenus à partir des répétitions de l'échantillon.

3.1 Production de données administratives

À partir des renseignements découlant des expériences du BNS, on définit quatre types de groupes ayant des probabilités connexes : groupe 1 : La personne n'est pas du tout représentée dans les données administratives et est supprimée; groupe 2 : La personne est déménagée dans un autre emplacement géographique et l'emplacement géographique a été modifié dans les données; groupe 3 : La personne est enregistrée dans les données administratives et figure au bon emplacement, ce qui fait en sorte qu'il n'y a aucun changement; groupe 4 : La personne a un enregistrement en double dans un autre emplacement géographique et figure donc également dans l'autre emplacement (doublon).

Pour chaque groupe, on définit une probabilité qui varie selon les strates définies par le sexe, le groupe ethnique (Blanc ou non-Blanc) et le groupe d'âge (moins de 30 ans, 31 à 44 ans, et 45 ans et plus), comme on peut le voir dans le tableau 3.1

Tableau 3.1
Probabilités selon les strates pour produire des données administratives

Groupe d'âge	Groupe ethnique	Sexe	p1	p2	p3	p4
Moins de 30 ans	Blanc	Hommes	0,45	0,15	0,3	0,1
31 à 44 ans	Blanc	Hommes	0,2	0,2	0,5	0,1
45 ans et plus	Blanc	Hommes	0,1	0,15	0,7	0,05
Moins de 30 ans	Blanc	Femmes	0,4	0,15	0,35	0,1
31 à 44 ans	Blanc	Femmes	0,15	0,15	0,55	0,15
45 ans et plus	Blanc	Femmes	0,1	0,1	0,75	0,05
Moins de 30 ans	Non-Blanc	Hommes	0,37	0,15	0,43	0,05
31 à 44 ans	Non-Blanc	Hommes	0,15	0,2	0,6	0,05
45 ans et plus	Non-Blanc	Hommes	0,05	0,13	0,8	0,02
Moins de 30 ans	Non-Blanc	Femmes	0,33	0,15	0,47	0,05
31 à 44 ans	Non-Blanc	Femmes	0,1	0,18	0,67	0,05
45 ans et plus	Non-Blanc	Femmes	0,05	0,1	0,83	0,02

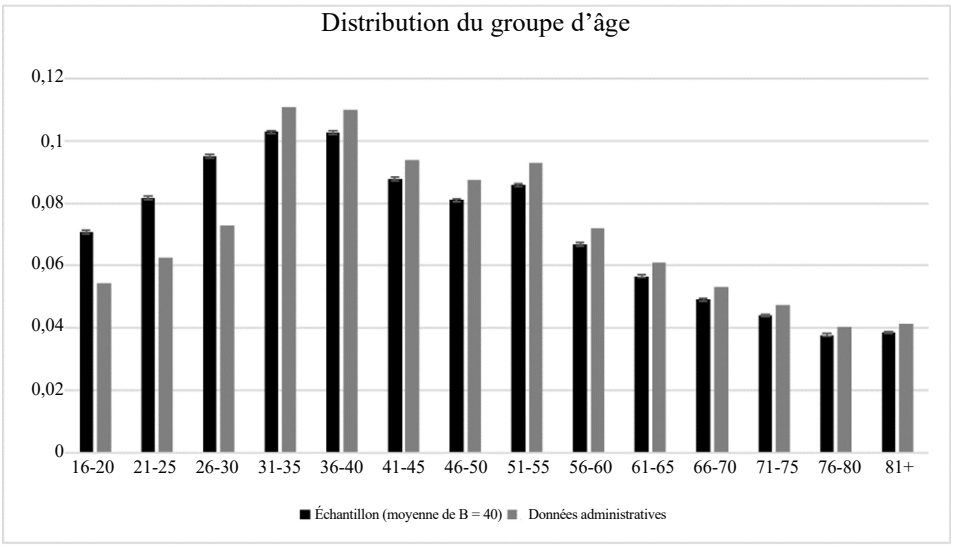
Note : p1 – personne supprimée des données administratives; p2 – personne ayant déménagée dans une autre région géographique; p3 – personne figurant dans les données administratives et au bon emplacement; p4 – personne figurant dans deux régions géographiques.

On sélectionne de manière aléatoire les personnes touchées dans chacune des strates et l'on prend les mesures appropriées dans chacun des groupes, afin de produire les données administratives générées. Après cette étape, on obtient un ensemble définitif de données administratives comptant $M = 1\ 026\ 969$ enregistrements (un ensemble plus petit que l'ensemble initial de données, qui comptait $N = 1\ 163\ 659$ enregistrements).

3.2 Résultats

Dans la figure 3.1, on compare deux distributions univariées et une bivariée de la moyenne des B = 40 chiffres pondérés d'enquête (et leurs intervalles de confiance simulés) et des données administratives décrites à la section 3.1.

Figure 3.1 Distributions a) du groupe d'âge, b) de l'élément géographique (autorité locale) et c) de l'élément géographique (autorité locale) × le sexe, comparant les chiffres pondérés moyens d'enquête (B = 40) et les données administratives (avec des intervalles de confiance simulés)



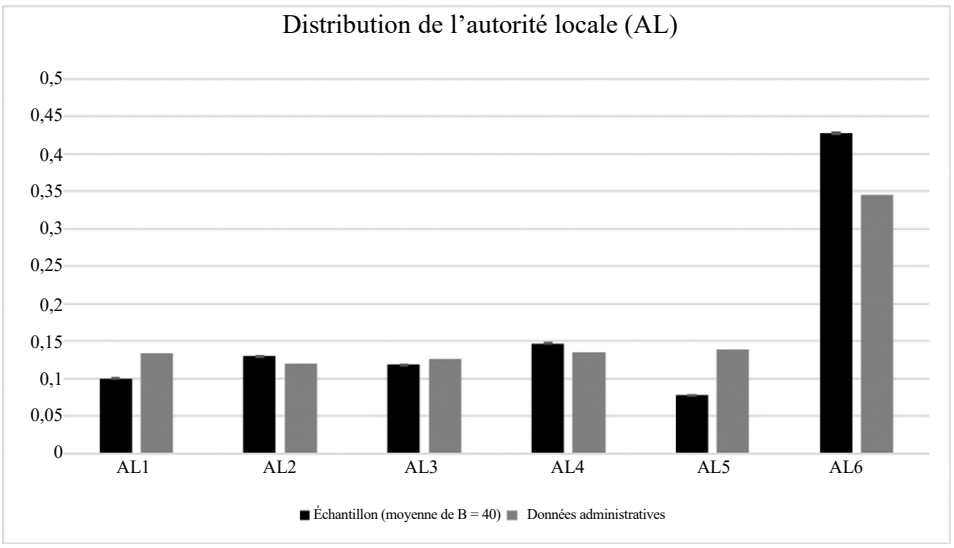
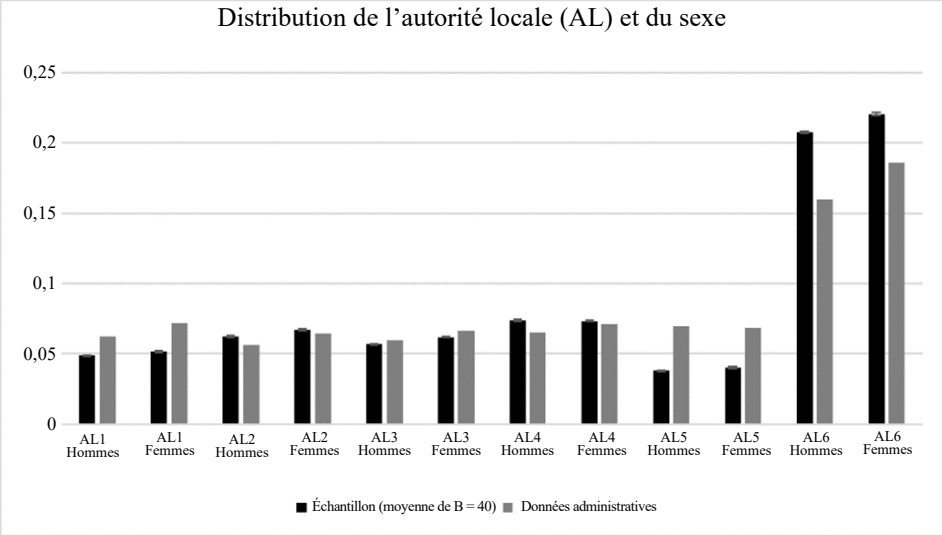


Figure 3.1 (suite) Distributions a) du groupe d'âge, b) de l'élément géographique (autorité locale) et c) de l'élément géographique (autorité locale) × le sexe, comparant les chiffres pondérés moyens d'enquête (B = 40) et les données administratives (avec des intervalles de confiance simulés)



L'approche utilisée pour produire les données administratives a été mise en application différemment dans le cas des groupes d'âge jeunes et des groupes d'âge plus vieux, car les jeunes ont tendance à quitter davantage la maison, par exemple pour loger dans une résidence universitaire. On peut donc observer, dans le premier tracé (a) de la figure 3.1, une sous-représentation parmi les groupes d'âge plus jeunes par rapport à ceux plus vieux. On a également sélectionné de manière arbitraire l'autorité locale (AL) 6 pour déplacer des personnes d'une AL vers une autre. C'est la raison pour laquelle on constate une grande divergence entre les chiffres des données administratives et les chiffres pondérés d'enquête pour l'AL6 dans le deuxième tracé (b) à la figure 3.1. Enfin, on montre les répercussions sur une distribution bivariée, comme on peut le voir dans le dernier tracé (c) de la figure 3.1, où les hommes étaient légèrement plus susceptibles que les femmes d'être associés au mauvais emplacement géographique AL6.

Passons maintenant aux indicateurs de qualité décrits à la section 2. Dans le tableau 3.2, on fournit les mesures de la distance $(1-\mathrm{ID})$ et $(1-\mathrm{DH})$ pour les distributions univariées et bivariées croisées avec les éléments géographiques.

Dans le tableau 3.2, on peut voir que les variables « Élément géographique » et « Groupe d'âge » comportent les mesures « $1-$distance » les plus petites par rapport aux autres distributions univariées. Ce résultat a une incidence sur toutes les distributions bivariées lorsque l'on utilise la variable « Élément géographique », surtout lorsqu'elle est croisée avec la variable « Groupe d'âge » comportant 14 catégories. Les mesures de la distance révèlent aussi que le nombre de catégories a une incidence sur ces mesures.

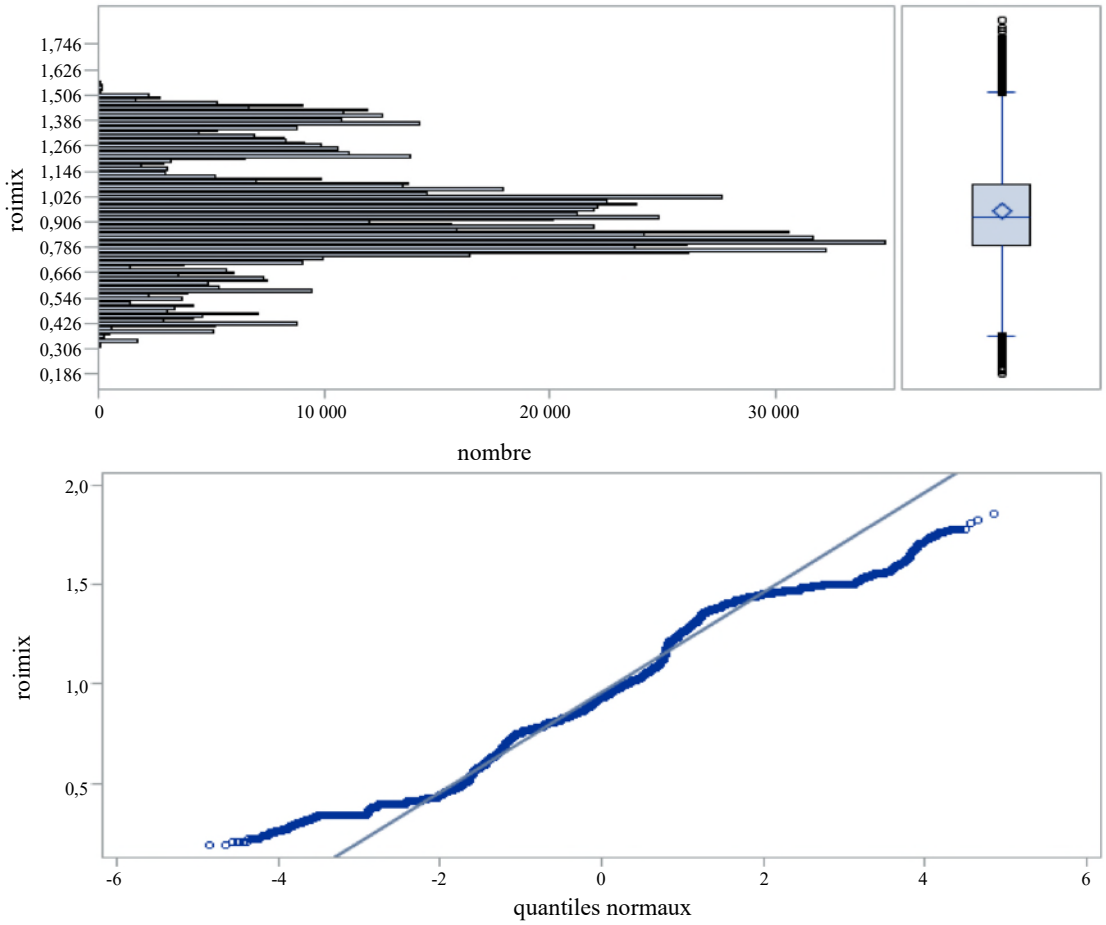
Maintenant, on se penche sur l'estimation des scores de propension et les calculs de l'indicateur R. La figure 3.2 montre l'histogramme des scores de propension estimés (étiquetés « roimix ») et le graphique de quantile-normale en fonction de la distribution normale pour le premier échantillon (B = 1). Les tracés sont semblables pour tous les échantillons.

Tableau 3.2
Indicateur de dissimilarité (1 - ID) et distance de Hellinger (1 - DH) (comme ils sont décrits à la section 2.1) pour un ensemble de distributions univariées et bivariées

Variable	Indicateur de dissimilarité (1 - ID) (erreur-type)	Distance de Hellinger (1 - DH) (erreur-type)
Élément géographique (6)	0,8973 (0,00049)	0,9103 (0,00043)
Groupe d'âge (14)	0,9424 (0,00047)	0,9500 (0,00038)
Groupe ethnique (16)	0,9955 (0,00018)	0,9909 (0,00025)
État matrimonial (6)	0,9603 (0,00049)	0,9688(0,00037)
Situation vis-à-vis de l'activité économique (10)	0,9732 (0,00047)	0,9738 (0,00032)
Élément géographique et sexe (12)	0,8972 (0,00048)	0,9091 (0,00043)
Élément géographique et groupe d'âge (84)	0,8836 (0,00049)	0,8900 (0,00041)
Élément géographique et groupe ethnique (96)	0,8900 (0,00047)	0,9034 (0,00042)
Élément géographique et état matrimonial (36)	0,8924 (0,00054)	0,9016 (0,00042)
Élément géographique et situation vis-à-vis de l'activité économique (60)	0,8927 (0,00048)	0,9014 (0,00042)

Note : Les erreurs-types de simulation sont présentées entre parenthèses.

Figure 3.2 Histogramme et graphique de quantile-normale des scores de propension estimés (étiquetés « roimix ») pour l'échantillon 1



Comme on peut le constater, dans la figure 3.2, on trouve des écarts par rapport à la distribution normale dans les « queues » comme on s'y attendrait lors de l'estimation des scores de propension à l'aide d'une régression linéaire, et il y a un bon ajustement au milieu de la distribution. On observe également des scores de propension supérieurs à 1 en raison de l'indicateur de représentativité d élevé des données administratives montrant les grandes propensions attendues à inclure dans les données administratives.

Dans le cas de l'indicateur R représentant la variation globale des scores de propension, on obtient les valeurs suivantes (l'erreur-type de simulation est présentée entre parenthèses) :

Type 1 (covariance de la population) : R = 0,7228 (0,0013).

Type 2 (mélange de données administratives et de moyennes démographiques) : R = 0,7413 (0,0010).

L'indicateur R est borné par 1 et (près de) 0. C'est la raison pour laquelle cette valeur est légèrement élevée. On peut donc affirmer en toute confiance que l'on a une représentativité globale supérieure à la moyenne des données administratives par rapport aux repères démographiques auxiliaires (d'échantillon pondéré).

Le tableau 3.3 renferme les indicateurs R partiels au niveau de la variable. Dans ce cas-ci, plus l'indicateur R partiel au niveau de la variable est élevé, plus il favorise le manque de représentativité.

Tableau 3.3
Indicateurs R partiels au niveau de la variable (erreurs-types de simulation)

Variable	Indicateur R partiel : type 1	Indicateur R partiel : type 2
Élément géographique (6)	0,2869 (0,00218)	0,2106 (0,00093)
Sexe (2)	0,0145 (0,00094)	0,0234 (0,00087)
Groupe d'âge (14)	0,1005 (0,00081)	0,1312 (0,00105)
Groupe ethnique (16)	0,0288 (0,00069)	0,0230 (0,00064)
État matrimonial (6)	0,0617 (0,00088)	0,0794 (0,00096)
Situation vis-à-vis de l'activité économique (10)	0,0534 (0,00068)	0,0674 (0,00086)

Dans le tableau 3.3, on peut voir que tous les indicateurs R partiels au niveau de la variable dans les calculs du type 1 et du type 2 diffèrent significativement de 0, car les intervalles de confiance n'incluent pas 0.

Les variables « Élément géographique » et « Groupe d'âge » ont des indicateurs R partiels au niveau de la variable plus élevés pour les calculs du type 1 et du type 2, et favorisent le plus le manque de représentativité. Cela était évident au chapitre des mesures de la distance dans le tableau 3.2, même si, dans le cas des indicateurs R partiels au niveau de la variable, on tient compte du manque de représentativité multivarié par rapport au calcul des mesures de la distance en fonction des distributions uniques. Les résultats vont de pair avec la façon dont on a produit les données administratives.

Ensuite, on évalue plus attentivement les variables « Élément géographique » et « Groupe d'âge » en ce qui concerne leurs catégories au moyen des indicateurs R partiels au niveau de la catégorie figurant dans les tableaux 3.4a et 3.4b, respectivement.

Dans le cas de la variable « Élément géographique » dans le tableau 3.4a, l'AL5 affiche la surreprésentation la plus élevée dans les données administratives pour les deux types de calculs, alors que l'AL6 affiche la sous-représentation la plus élevée. En ce qui concerne la variable « Groupe d'âge » dans le tableau 3.4b, les groupes d'âge plus jeunes affichent une sous-représentation élevée par rapport aux groupes d'âge plus vieux pour les deux types.

Après, dans le tableau 3.5, on montre les résultats de la mise en œuvre de l'approche bootstrap à deux étapes vue à la section 2.3, afin d'estimer les erreurs-types pour l'indicateur R et les indicateurs R partiels au niveau de la variable (tableau 3.3). On a effectué un rééchantillonnage avec remise pour l'échantillon et les données administratives pour B = 60, et l'on a calculé les scores de propension, les indicateurs R ainsi que les erreurs-types découlant des indicateurs R. Le tableau 3.5 montre certaines légères différences entre les erreurs-types de simulation et les erreurs-types découlant de l'approche bootstrap, même si elles ont une ampleur semblable.

Tableau 3.4a
Indicateurs R partiels au niveau de la catégorie pour la variable « Élément géographique » (erreurs-types de simulation)

Élément géographique	Indicateur R partiel au niveau de la catégorie : type 1	Indicateur R partiel au niveau de la catégorie : type 2
AL1	0,0858 (0,00122)	0,0821 (0,00072)
AL2	-0,0441 (0,00080)	-0,0241 (0,00081)
AL3	-0,0065 (0,00096)	0,0200 (0,00079)
AL4	-0,0464 (0,00076)	-0,0272 (0,00078)
AL5	0,2270 (0,00231)	0,1437 (0,00081)
AL6	-0,1382 (0,00089)	-0,1229 (0,00076)

Note : AL = autorité locale.

Tableau 3.4b
Indicateurs R partiels au niveau de la catégorie pour la variable « Groupe d'âge » (erreurs-types de simulation)

Groupe d'âge	Indicateur R partiel au niveau de la catégorie : type 1	Indicateur R partiel au niveau de la catégorie : type 2
16 à 20 ans	-0,0455 (0,00070)	-0,0619 (0,00110)
21 à 25 ans	-0,0520 (0,00073)	-0,0675 (0,00106)
26 à 30 ans	-0,0545 (0,00073)	-0,0723 (0,00107)
31 à 35 ans	0,0211 (0,00093)	0,0216 (0,00078)
36 à 40 ans	0,0187 (0,00099)	0,0194 (0,00081)
41 à 45 ans	0,0180 (0,00092)	0,0173 (0,00078)
46 à 50 ans	0,0105 (0,00100)	0,0190 (0,00086)
51 à 55 ans	0,0127 (0,00085)	0,0207 (0,00075)
56 à 60 ans	0,0114 (0,00113)	0,0170 (0,00092)
61 à 65 ans	0,0113 (0,00116)	0,0158 (0,00097)
66 à 70 ans	0,0108 (0,00102)	0,0157 (0,00087)
71 à 75 ans	0,0089 (0,00074)	0,0139 (0,00059)
76 à 80 ans	0,0068 (0,00122)	0,0124 (0,00102)
81 ans et plus	0,0044 (0,00084)	0,0116 (0,00066)

Tableau 3.5
Comparaison des erreurs-types estimées à partir de l'approche bootstrap à deux étapes aux erreurs-types de simulation pour les indicateurs R et les indicateurs R partiels au niveau de la variable

Variable	Indicateur R partiel : type 1		Indicateur R partiel : type 2	
	Erreur-type de simulation	Erreur-type estimée à partir de l'approche bootstrap à deux étapes	Erreur-type de simulation	Erreur-type estimée à partir de l'approche bootstrap à deux étapes
Indicateur R	0,0013	0,0010	0,0010	0,0007
Élément géographique (6)	0,00218	0,00145	0,00093	0,00064
Sexe (2)	0,00094	0,00082	0,00087	0,00082
Groupe d'âge (14)	0,00081	0,00060	0,00105	0,00087
Groupe ethnique (16)	0,00069	0,00066	0,00064	0,00060
État matrimonial (6)	0,00088	0,00062	0,00096	0,00077
Situation vis-à-vis de l'activité économique (10)	0,00068	0,00064	0,00086	0,00090

En résumé, les résultats concernant l'indicateur R et ses indicateurs R partiels correspondent aux attentes en raison de la façon dont on a produit les données administratives pour l'étude par simulation. On montre, dans le cas présent, les deux types de calculs de l'indicateur R et des indicateurs R partiels selon l'information auxiliaire accessible pour les variables d'intérêt : le type 1, qui comprend des chiffres bivariés complets connus, ou le type 2, qui comprend uniquement des chiffres marginaux connus. Les indicateurs R partiels au niveau de la catégorie montrent des erreurs-types plus élevées pour l'information bivariée complète dans les calculs du type 1 par rapport aux erreurs-types obtenues lorsque l'on n'utilise que l'information marginale dans les calculs du type 2, même si ce n'était pas évident pour les indicateurs R partiels au niveau de la variable.

Selon tous les indicateurs R partiels au niveau de la catégorie, on peut établir le profil des personnes qui sont sous-représentées ou surreprésentées dans les données administratives en tenant compte de la nature multivariée du manque de représentativité. Par exemple, on constate que les groupes d'âge plus jeunes de l'AL6 sont sous-représentés. Il est important de chercher d'autres sources de données administratives qui peuvent permettre de corriger cette sous-représentation.

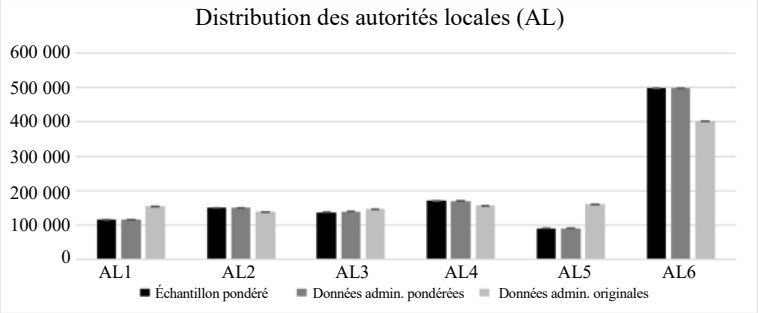
3.3 Tenir compte du manque de représentativité

Pour calculer les indicateurs R, on a estimé les scores de propension associés au fait d'être représenté dans les données administratives. On peut maintenant utiliser l'inverse de ces scores de propension pour pondérer les données administratives et améliorer l'estimation des distributions dans les données administratives. On souligne que, puisque l'on a utilisé un modèle de régression linéaire pour estimer les scores de propension, le poids est semblable au « poids g » dans les estimateurs standard par la régression généralisée

(Särndal et Lundström, 2008). Dans la figure 3.3, on montre les distributions de certaines variables des chiffres pondérés d'échantillon, les données administratives pondérées par l'inverse de la propension selon le calcul du type 1 et les données administratives originales pour lesquelles on utilise un facteur de correction unique selon le poids de représentativité d afin de veiller à ce que le total des données administratives soit égal à la taille de l'échantillon pondéré : facteur de correction C = 1163650/1026969 = 1,133. Sous chaque figure, on fournit également les mesures de la distance $(1-\mathrm{DH})$ comparant 1) la distribution de l'échantillon pondéré à la distribution des données administratives pondérées par l'inverse de la propension et 2) la distribution de l'échantillon pondéré à la distribution des données administratives originales.

En résumé, on peut constater que le fait de pondérer les données administratives par l'inverse des scores de propension améliore l'estimation des distributions par rapport à l'utilisation des données administratives originales. Cela est particulièrement vrai dans le cas des variables utilisées pour produire les données administratives simulées, pour lesquelles il y avait de grandes différences entre les données administratives originales et les chiffres pondérés de l'échantillon, comme les variables « Groupe d'âge » et « Élément géographique ». En outre, la pondération permet de corriger l'estimation des distributions dans les données administratives dans le cas des variables qui n'ont pas été utilisées pour la production des données administratives générées. Par exemple, la variable « Statut économique » n'a pas été utilisée pour produire les données administratives. Malgré tout, la mesure $(1-\mathrm{DH})$ a été améliorée lorsque l'on utilisait la pondération par l'inverse de la propension par rapport aux données administratives originales : $(1-\mathrm{DH}) = 0{,}9962$ à $(1-\mathrm{DH}) = 0{,}9738$, respectivement. Il convient de souligner que les variables présentées dans la figure 3.3 étaient comprises dans le modèle pour estimer les scores de propension. Il est donc important de tenir compte de toutes les variables d'intérêt lorsque l'on estime les scores de propension.

Figure 3.3 Comparaison des chiffres pondérés de l'échantillon, des données administratives pondérées par l'inverse de la propension (type 1) et des données administratives originales pour les variables

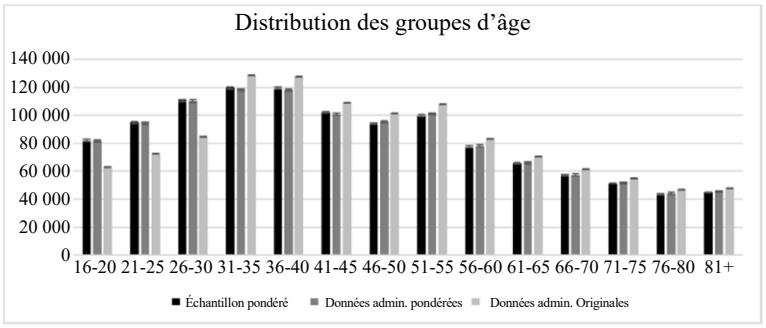


(1 - DH) pour échantillon pondéré, données administratives pondérées = 0,9981 (0,00005)
(1 - DH) pour échantillon pondéré, données administratives originales = 0,9103 (0,00043)

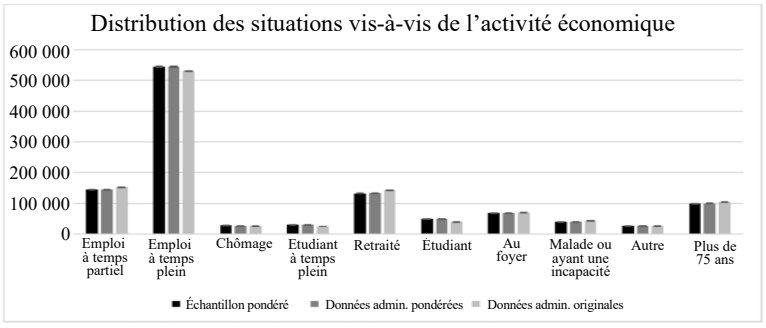
Notes : - a) élément géographique, b) groupe d'âge, c) situation vis-à-vis de l'activité économique et d) élément géographique × sexe (sous chaque figure on trouve les mesures de distance [1 - DH] pour [1] l'échantillon pondéré par rapport aux données administratives pondérées et [2] l'échantillon pondéré par rapport aux données administratives originales [l'erreur-type de simulation se trouve entre parenthèses]).

- DH = distance de Hellinger.

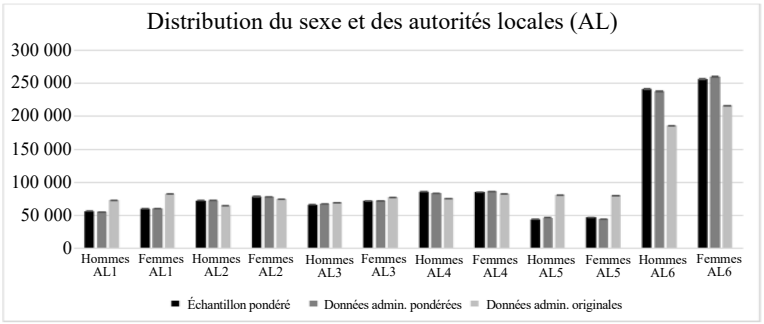
Figure 3.3 (suite) Comparaison des chiffres pondérés de l'échantillon, des données administratives pondérées par l'inverse de la propension (type 1) et des données administratives originales pour les variables



(1 - DH) pour échantillon pondéré, données administratives pondérées = 0,9956 (0,00010)
(1 - DH) pour échantillon pondéré, données administratives originales = 0,9500 (0,00038)



(1 - DH) pour échantillon pondéré, données administratives pondérées = 0,9962 (0,00010)
(1 - DH) pour échantillon pondéré, données administratives originales = 0,9738 (0,00032)



(1 - DH) pour échantillon pondéré, données administratives pondérées = 0,9911 (0,00034)
(1 - DH) pour échantillon pondéré, données administratives originales = 0,9091 (0,00043)

Notes : - a) élément géographique, b) groupe d'âge, c) situation vis-à-vis de l'activité économique et d) élément géographique × sexe (sous chaque figure on trouve les mesures de distance [1 - DH] pour [1] l'échantillon pondéré par rapport aux données administratives pondérées et [2] l'échantillon pondéré par rapport aux données administratives originales [l'erreur-type de simulation se trouve entre parenthèses]).

- DH = distance de Hellinger.

4. Application

En collaborant avec l'équipe du BNS, on a produit les indicateurs de qualité suggérés à l'aide du code R accessible sur GitHub (Kim et Shlomo, 2023) et en fonction du calcul du type 2. On cherchait à évaluer la représentativité d'une source combinée de données administratives s'intitulant Admin-Based Housing by Ethnicity Dataset (ABHED). Cet ensemble de données est hautement confidentiel. Il est interdit de l'utiliser

hors des serveurs sécurisés du BNS. On a donc collaboré pour produire les indicateurs de qualité suggérés, et l'équipe du BNS a mis en application le code R dans son environnement sécurisé. Les données administratives de l'ABHED ont été élaborées en jumelant trois sources distinctes de données administratives composites : la version 3.0 de l'ensemble de données administratives sur l'ethnicité, la version 3.0 des estimations administratives des ménages et la version 1.0 des données administratives sur le parc de logements. Dans cette application réalisée sur des serveurs sécurisés du BNS, l'équipe a réussi à utiliser les données du recensement de 2021 de l'Angleterre et du pays de Galles comme données démographiques auxiliaires comparatives. L'application était axée sur quatre variables figurant dans l'ensemble de données ABHED et dans l'ensemble de données du recensement de 2021, soit le sexe, le groupe d'âge, le type de logement et l'ethnicité, et a été réalisée pour une autorité locale. On trouve tous les renseignements sur les travaux réalisés pour évaluer la qualité des données administratives dans le rapport produit par le BNS (2023).

Dans le cas des distributions univariées et bivariées que l'on a comparées directement aux données démographiques auxiliaires à l'aide de mesures de la distance, on a découvert que l'ethnicité et le type de logement s'éloignaient le plus par rapport aux autres variables au niveau univarié selon la DH. En outre, cela a permis d'obtenir des mesures de la distance plus faibles lorsque l'on examinait les distributions bivariées avec l'ethnicité et le type de logement.

En ce qui concerne l'évaluation de l'indicateur R, l'indicateur R global était de 0,7565, ce qui correspond à une bonne représentativité dans son ensemble. Ensuite, on a calculé les indicateurs R partiels pour vérifier quelles variables, et quelles catégories, favorisaient le manque de représentativité. L'ethnicité était associée à l'indicateur R partiel au niveau de la variable le plus important ($R_U(x_k) = 0{,}0346$), et la catégorie « Autre » affichait un indicateur R partiel au niveau de la catégorie important, soit $R_U(x_k^h) = -0{,}1764$, ce qui montre que ce groupe est sous-représenté. Cette situation peut être attribuable à différentes définitions entre la source de données administratives et le repère démographique auxiliaire du recensement de 2021.

Le type de logement affichait aussi un indicateur R partiel au niveau de la variable considérable ($R_U(x_k) = 0{,}0172$), et la catégorie « Logement mitoyen » affichait une grande surreprésentation ($R_U(x_k^h) = 0{,}0798$), tout comme la catégorie « Immeuble commercial » ($R_U(x_k^h) = 0{,}0558$). La catégorie « Plain-pied construit expressément pour la location » affichait une grande sous-représentation ($R_U(x_k^h) = -0{,}0643$), tout comme la catégorie « Logement jumelé » ($R_U(x_k^h) = -0{,}0487$). La répartition plus uniforme des indicateurs R partiels au niveau de la catégorie pour le type de logement a permis de garantir que l'indicateur partiel au niveau de la variable « Logement » était plus petit que celui de la variable « Ethnicité » et favorisait donc moins le manque de représentativité. On trouve tous les renseignements sur les indicateurs R partiels dans le tableau 4.1, qui a été reproduit à partir du rapport du BNS (2023).

Il convient de souligner que les indicateurs R représentent le manque de représentativité multivarié par rapport aux mesures de distance et fournissent davantage d'information sur l'incidence de ces variables lorsqu'on les combine.

Tableau 4.1
Indicateurs R partiels au niveau de la variable et de la catégorie pour les données administratives du bureau national de la statistique : ensemble de données sur le logement en fonction de l'ethnicité (ABHED)

Variable	Indicateur R partiel
Sexe	0,0005
1. Femmes	0,0161
2. Hommes	-0,0167
Ethnicité	0,0346
1. Blanc	0,0364
2. Mixte	0,0038
3. Asiatique	-0,0444
4. Noir	-0,0118
5. Autre	-0,1764
Type de logement	0,0172
1. Logement isolé	0,0041
2. Logement jumelé	-0,0487
3. Logement mitoyen	0,0798
4. Logement dans un bloc de plains-pieds construit expressément pour la location ou immeuble à usage locatif	-0,0643
5. Logement faisant partie d'une maison transformée ou partagée, y compris les chambres meublées	-0,0265
6. Logement faisant partie d'un autre immeuble transformé, comme une ancienne école, une ancienne église ou un ancien entrepôt	0,0217
7. Logement faisant partie d'un immeuble commercial, comme un immeuble de bureaux, un hôtel ou un immeuble où est situé un magasin	0,0558
8. Roulotte ou autre structure mobile ou temporaire	-0,0012
Groupe d'âge	0,0059
1. 0 à 15 ans	0,0470
2. 16 à 24 ans	0,0241
3. 25 à 34 ans	-0,0392
4. 35 à 44 ans	-0,0344
5. 45 à 54 ans	-0,0119
6. 55 à -64 ans	-0,0065
7. 65 à 74 ans	0,0002
8. 75 ans et plus	0,0155

Note : Reproduit à partir du rapport du bureau national de la statistique (2023).

L'application ne comprenait pas l'estimation de la variance des indicateurs de qualité. Cependant, puisque le BNS a été en mesure d'utiliser les chiffres du recensement de 2021 comme repères démographiques pour estimer les indicateurs R, on prévoit de petites erreurs-types, comme on l'a constaté dans l'étude par simulation présentée à la section 3. Dans le cadre des travaux futurs, on ajoutera l'approche relative à la variance bootstrap à deux étapes au paquet sur GitHub (Kim et Shlomo, 2023).

En résumé, les résultats de cette application réelle montrent que l'évaluation de la qualité visant à comprendre le manque de représentativité dans les données administratives de l'ABHED a permis de désigner les variables et leurs catégories qui étaient moins représentatives par rapport aux repères démographiques généraux obtenus à l'aide du recensement de 2021.

5. Discussion

Il est important d'évaluer la représentativité des données administratives avant de les utiliser dans les systèmes statistiques nationaux. Les indicateurs de qualité suggérés sont motivés par le fait que les ONS

devraient exploiter davantage leurs données d'enquête de grande qualité, pour lesquelles on utilise des pratiques exemplaires afin de garantir une collecte de données fiable et l'on calcule des poids de sondage pour atténuer les biais de non-réponse et procéder à une calibration par rapport aux totaux démographiques connus. Les données d'enquête de grande qualité peuvent être utilisées pour évaluer la qualité des données administratives, surtout lorsqu'il est impossible d'avoir accès à des données de recensement. Dans le présent document, les indicateurs de qualité suggérés devraient aider à atteindre l'objectif visant à réaliser une évaluation quantitative de la représentativité des données administratives, car ils mettent en évidence les variables et leurs catégories qui favorisent le manque de représentativité. On a montré que les indicateurs de qualité obtenus permettent de bien évaluer la représentativité, mais dépendent des variables qui figurent dans le modèle pour estimer les scores de propension. De plus, on a montré que le fait d'utiliser l'inverse du score de propension comme poids dans les données administratives améliore l'estimation des distributions.

Par ailleurs, à la section 2.3, on a proposé une approche bootstrap à deux étapes pour estimer les erreurs-types des indicateurs et des estimations découlant de la PIP. On a présenté l'approche dans l'étude par simulation fournie à la section 3. L'étude par simulation a montré que les erreurs-types basées sur les échantillons aléatoires répétés étaient petites (bien que l'on n'ait pas tenu compte de la non-réponse dans les échantillons tirés, ce qui peut avoir augmenté la variation au chapitre des données repères de l'échantillon) et que les erreurs-types bootstrap étaient semblables aux erreurs-types de simulation dans le cas des indicateurs R. On a conclu que les indicateurs de qualité fournissent une indication fiable de la représentativité des données administratives en fonction des covariables figurant dans le modèle de score de propension.

Il est possible d'obtenir le code R et un guide de l'utilisateur sur GitHub (Kim et Shlomo, 2023). Le paquet R prend comme données d'entrée l'ensemble de données administratives et un échantillon probabiliste de référence ayant des poids de sondage appropriés, afin de produire les indicateurs de qualité, comme le montre le présent document.

En consultation avec le BNS, pour tenir compte des très grands ensembles de données administratives dont il dispose (certains comptant plus de 50 millions de personnes), on a élaboré un nouveau code R sur la plateforme GitHub pour estimer les scores de propension. Le nouveau code R permet d'estimer les scores de propension à partir des tableaux de fréquences croisés par les variables du modèle de score de propension plutôt qu'à partir des microdonnées individuelles. Cela évite d'avoir à estimer les scores de propension à partir des microdonnées et réduit considérablement les ressources de calcul nécessaires pour calculer les indicateurs de qualité.

Remerciements

La présente étude a été financée par la subvention suivante : « Methodological Advancements on the Use of Administrative Data in Official Statistics », Economic and Social Research Council, ES/V005456/1, Royaume-Uni.

Bibliographie

Alho, J.M. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46(3), 623-635.

Austin, P.C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3), 399-424.

Bianchi, A., Shlomo, N., Schouten, B., Da Silva, D.N. et Skinner, C. (2019). Estimation des propensions à répondre et indicateurs de représentativité des réponses utilisant l'information au niveau de la population. *Techniques d'enquête*, 45(2), 231-264. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2019002/article/00009-fra.pdf.

Bureau national de la statistique (BNS) (2016). *Patient Register: Quality Assurance of Administrative Data used in Population Statistics*. Office for National Statistics, Royaume-Uni, disponible à l'adresse : https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/patientregisterqualityassuranceofadministrativedatausedinpopulationstatisticsdec2016 [consulté le 25/02/2024].

Bureau national de la statistique (BNS) (2022). *Coverage Estimation for Census 2021 in England and Wales*. Site web de l'ONS, article méthodologique, (publié le 9 novembre 2022). Disponible à l'adresse : https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/coverageestimationforcensus2021inenglandandwales [consulté le 11/07/2024].

Bureau national de la statistique (BNS) (2023). *Quality Indicators for Representativeness in Administrative Data: R-indicators and Distance Metrics*. Office for National Statistics, Royaume-Uni. Disponible à l'adresse : https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/qualityindicatorsforrepresentativenessinadministrativedatarindicatorsanddistancemetrics#:~:text=R%2Dindicators%20and%20distance%20metrics%20use%20variables%20from%20the%20census,auxiliary%20data%20in%20this%20paper [consulté le 25/02/2024].

Chen, Y., Li, P., Rao, J.N.K. et Wu, C. (2022). Pseudo empirical likelihood inference for nonprobability survey samples. *The Canadian Journal of Statistics*, 50(4), 1166-1185.

Chen, Y., Li, P. et Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statstistical Association*, 115(532), 2011-2021.

Commission économique des Nations Unies pour l'Europe (CEE-ONU) (2021). *Guidelines for Assessing the Quality of Administrative Sources for Use in Censuses*. Conference of European Statisticians Task Force, Nations Unies, Genève, Suisse. Disponible à l'adresse : https://unece.org/sites/default/files/2021-10/ECECESSTAT20214_WEB.pdf [consulté le 25/02/2024].

Daas, P., Ossen, S., Vis-Visschers, R. et Arends-Toth, J. (2009). *Checklist for the Quality Evaluation of Administrative Data Sources*. Document de discussion (09042), Statistics Netherlands, La Haye, Pays-Bas, disponible sur : https://ec.europa.eu/eurostat/documents/64157/4374310/45-Checklist-quality-evaluationadministrative-data-sources-2009.pdf/24ffb3dd-5509-4f7e-9683-4477be82ee60 [consulté le 25/02/2024].

Duncan, O.D. et Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, 20(2), 210-217. https://doi.org/10.2307/2088328.

Dunstan, S. et Correia, S. (2021). *Framework for Assuring the Quality of Administrative Data for Use in the 2021 Census of England and Wales*. Document de travail EAP144, UK Statistics Authority. Royaume-Uni, disponible à l'adresse : https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2021/05/EAP144-Framework-quality-assurance-of-admin-data.pdf [consulté le 25/02/2024].

Eurostat (2019). *Quality Assurance Framework of the European Statistical System, Version 2*. Disponible sur : https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646 [consulté le 14/04/2025].

Kim, J.-K. et Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 24.1, 375-394.

Kim, M.S. et Shlomo, N. (2023). *Quality Indicators for Administrative Data User Manual for R Package on GitHub*. Disponible sur : https://github.com/sook-tusk/qualadmin [consulté le 25/02/2024].

Nations Unies (NU) (2019). *Les Nations Unies Manuel des cadres nationaux d'assurance qualité pour les statistiques officielles notamment les recommandations, le cadre et les directives de mise en oeuvre*. Département des affaires économiques et sociales, Série d'études méthodologiques ST/ESA/STAT/SER.M/100. Disponible à l'adresse https://unstats.un.org/unsd/methodology/dataquality/references/1902216-UNNQAFManual-WEB.pdf [consulté le 25/02/2024].

Opsomer, J.D. et Erciulescu, A.L. (2021). Replication variance estimation after sample-based calibration. *Techniques d'enquête*, 47(2), 265-277. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021002/article/00006-fra.pdf.

Ouwehand, P. et Schouten, B. (2014). Measuring representativeness of short-term business statistics. *Journal of Official Statistics*, 30(4), 623-649.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.

Särndal, C.-E. et Lundström, S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, 24(2), 167-191.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35(1), 107-121. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2009001/article/10887-fra.pdf.

Schouten, B. et Shlomo, N. (2017). Selecting adaptive survey design strata with partial R-indicators. *Revue Internationale de Statistique*, 85(1), 143-163.

Shlomo, N., Luiten, A. et Schouten, B. (2022). Representativeness of 2011 EU-SILC responses and response rates over time. Dans *Improving the Measurement of Poverty and Social Exclusion in Europe: Reducing Nonsampling Errors*, (Éds., P. Lynn et L. Lyberg), Publications Office of the European Union. Disponible sur : https://op.europa.eu/en/publication-detail/-/publication/798e3ef9-fe65-11ec-b94a-01aa75ed71a1/language-en [consulté le 15/04/2025].

Spector, P.E. (1992). *Summated Rating Scale Construction*. Sage.

Wu, C. (2022). Inférence statistique avec des échantillons d'enquête non probabiliste. *Techniques d'enquête*, 48(2), 307-338. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2022002/article/00002-fra.pdf.

Zhang, L. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66, 41-63.