No 75-005-M au catalogue
ISSN 2292-3799
ISBN 978-0-660-73232-9

Statistiques sur le travail : Documents techniques

La méthodologie d'estimation sur petits domaines en utilisant les données de l'Enquête sur la population active

Par François Verret, Braedan Walker, Cynthia Bocci et Jean-François Beaumont

Date de diffusion : le 17 septembre 2024

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.

La méthodologie d'estimation sur petits domaines en utilisant les données de l'Enquête sur la population active

Par **François Verret, Braedan Walker, Cynthia Bocci** et **Jean-François Beaumont**

L'Enquête sur la population active (EPA) de Statistique Canada produit des estimations des caractéristiques de la population active pour différents niveaux géographiques au Canada. Bien que des estimations pondérées standard, c.-à-d. des estimations directes, puissent être produites à partir de l'EPA, ces estimations sont moins précises pour de nombreuses petites collectivités et régions rurales en raison de la petite taille de l'échantillon dans ces régions.

Dans le cadre du plan d'action sur les données désagrégées, une méthode d'estimation sur petits domaines (EPD) a été mise au point pour produire des estimations mensuelles des caractéristiques de la population active pour des niveaux géographiques plus détaillés que ceux disponibles directement à partir de l'EPA, à savoir les régions métropolitaines de recensement (RMR), les agglomérations de recensement (AR), et une géographie complémentaire appelée « zones de travail autonomes » (ZTA). Les ZTA sont des zones fonctionnelles composées de subdivisions de recensement regroupées en fonction des tendances du navettage (OCDE, 2020).

1. Introduction

Dans l'Enquête sur la population active (EPA), des estimations directes pour une région donnée sont obtenues à l'aide de données d'enquête provenant de cette région ainsi que d'un ensemble de poids d'enquête appropriés. Des estimations directes sont ainsi fiables si la taille de l'échantillon pour la région d'intérêt est suffisamment grande.

Les méthodes d'estimation sur petits domaines (EPD) tentent de produire des estimations fiables lorsque la taille de l'échantillon d'une région est faible. Pour ce faire, la petite quantité de données d'enquête est complémentée par des renseignements supplémentaires qui prennent la forme d'un modèle utilisant des données auxiliaires. Les estimations sur petits domaines des chiffres de l'emploi et des taux de chômage ont été produites à l'aide d'un modèle au niveau de la région. Un modèle au niveau de la région est un ensemble d'hypothèses sur la relation entre les estimations directes et les données auxiliaires pour la région. Dans la présente application, nous avons considéré un modèle linéaire à erreurs normales. L'estimation sur petits domaines obtenue pour une région donnée est une combinaison linéaire de l'estimation directe pour cette région et d'une prédiction découlant du modèle. Cette dernière est souvent appelée estimation synthétique et comprend des données d'enquête provenant de la région d'intérêt ainsi que d'autres régions utilisées dans la modélisation. Ainsi, les estimations sur petits domaines sont des estimations indirectes en ce sens qu'elles sont fondées sur des données d'enquête hors de la région d'intérêt. Pour les régions dont les échantillons sont les plus petits, les estimations directes ne sont pas fiables et les estimations sur petits domaines sont principalement basées sur les prédictions du modèle. Cependant, pour les régions avec les tailles d'échantillon les plus grandes, c'est l'inverse et les estimations sur petits domaines tendent à être proches des estimations directes.

Comme indiqué ci-dessus, l'utilisation d'un modèle au niveau de la région nécessite des données auxiliaires disponibles pour la région. Il n'est pas nécessaire de disposer de ces renseignements au niveau de l'unité (la personne). Toutefois, elles doivent provenir d'une source indépendante du mécanisme de sélection de l'échantillon de l'enquête. Dans l'EPA, le nombre de bénéficiaires de l'assurance-emploi dans chaque région d'intérêt a été utilisé ainsi que les projections démographiques du nombre de personnes âgées de 15 à 64 ans et de 65 ans et plus. Le premier nombre est obtenu à partir d'une source administrative et les autres nombres sont obtenus auprès du Centre de démographie de Statistique Canada. La production d'estimations sur petits domaines dépend de la disponibilité de ces données auxiliaires pour la période de référence et, de ce fait, les estimations sur petits domaines sont moins actuelles que les estimations directes.

La section 2 du présent document décrit le modèle au niveau de la région utilisé pour produire les chiffres de l'emploi et des taux de chômage pour les RMR, les AR et les ZTA. Dans la section 3, l'évaluation des estimations sur petits domaines et les diagnostics utilisés pour la validation du modèle sont brièvement discutés.

Les estimations sur petits domaines sont obtenues indépendamment chaque mois en utilisant le module d'estimation sur petits domaines du logiciel généralisé G-EST version 2.03 (Estevao et coll., 2023a, 2023b).

2. Modèle au niveau de la région

Supposons que nous souhaitons estimer un certain paramètre de population pour une région donnée i et un mois donné, que nous pouvons désigner par θ_i. Par exemple, θ_i pourrait être le nombre de personnes avec emploi ou le taux de chômage de la région i. Un estimateur direct de θ_i est désigné par $\hat{\theta}_i$. L'erreur d'échantillonnage est représentée par l'expression $e_i = \hat{\theta}_i - \theta_i$. Le modèle de Fay-Herriot est le modèle au niveau de la région le plus courant. Il comprend deux composantes : le modèle d'échantillonnage et le modèle de liaison.

Le modèle d'échantillonnage peut être exprimé comme suit :

$$\hat{\theta}_i = \theta_i + e_i ,$$

où $E_p(e_i) = 0$ et $\text{var}_p(e_i) = \psi_i$. L'indice p indique que l'espérance et la variance se rapportent au plan d'échantillonnage (ou au mécanisme de sélection de l'échantillon). L'hypothèse implicite est que l'estimateur direct ne présente pas de biais dans le cadre du plan d'échantillonnage. Cette hypothèse semble raisonnable dans le cas de l'EPA, étant donné les types de paramètres pris en compte et les méthodes d'estimation utilisées. La quantité ψ_i représente la variance de $\hat{\theta}_i$ relativement au plan d'échantillonnage et est généralement inconnue. Un estimateur direct de ψ_i est désigné par $\hat{\psi}_i$. Dans le cadre de l'EPA, $\hat{\psi}_i$ est obtenu au moyen d'une méthode de bootstrap.

La deuxième composante du modèle de Fay-Herriot est le modèle de liaison. Un modèle de liaison courant est :

$$\theta_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i ,$$

où $E_m(v_i|\mathbf{z}_i) = 0$, $\text{var}_m(v_i|\mathbf{z}_i) = \sigma_v^2$, $\mathbf{z}_i$ est un vecteur de variables auxiliaires pour la région i, alors que $\boldsymbol{\beta}$ et σ_v^2 sont des paramètres inconnus du modèle. La définition de $\mathbf{z}_i$ pour le nombre de personnes avec emploi et le taux de chômage est fournie à la fin de cette section. L'indice m indique que l'espérance et la variance se rapportent au modèle.

Outre les hypothèses du modèle ci-dessus, on suppose généralement que les erreurs e_i et v_i, $i = 1,...,M$ ont une distribution normale et sont mutuellement indépendantes. La quantité M est le nombre de régions (RMR, AR, ZTA) utilisées pour la modélisation.

En combinant les modèles d'échantillonnage et de liaison, nous obtenons le modèle de Fay-Herriot :

$$\hat{\theta}_i = \mathbf{z}_i'\boldsymbol{\beta} + a_i ,$$

où $a_i = v_i + e_i$, $E_{mp}(a_i|\ _i) = 0$, $\text{var}_{mp}(a_i|\mathbf{z}_i) = \sigma_v^2 + \tilde{\psi}_i$ et $\tilde{\psi}_i = E_m(\psi_i|\mathbf{z}_i)$ est une variance lissée. L'indice mp indique que l'espérance et la variance se rapportent au modèle ainsi qu'au plan d'échantillonnage. En supposant que $\tilde{\psi}_i$ est connue, les estimations des paramètres du modèle $\boldsymbol{\beta}$ et σ_v^2 peuvent être obtenues à l'aide de la méthode du maximum de vraisemblance restreint et sont désignées par $\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}_v^2$.

L'estimation synthétique de θ_i est simplement la valeur prédite $\mathbf{z}_i'\hat{\boldsymbol{\beta}}$. On l'appelle une estimation indirecte, car $\hat{\boldsymbol{\beta}}$ est obtenu à l'aide de l'estimation directe de la région i, $\hat{\theta}_i$, ainsi que des estimations directes des régions autres que i. Parce qu'elle a recours à davantage de données, l'estimation synthétique est généralement plus stable que l'estimation directe, en particulier pour les régions à taille d'échantillon réduite. Cependant, l'estimation synthétique

se fie à la formulation correcte du modèle de Fay-Herriot et peut faire l'objet d'un biais important, en cas d'erreur de spécification dans le modèle de Fay-Herriot. Un estimateur plus robuste et encore plus stable utilisé dans le cadre de l'EPA est l'estimateur composite

$$\hat{\theta}_i^{EPD} = \hat{\gamma}_i \hat{\theta}_i + (1-\hat{\gamma}_i)\mathbf{z}_i'\hat{\boldsymbol{\beta}},$$

où $\hat{\gamma}_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \tilde{\psi}_i)$. Cette estimation sur petits domaines est une moyenne pondérée de l'estimation directe et de l'estimation synthétique. Lorsque l'estimation directe est stable, $\tilde{\psi}_i$ tend à être petite et l'estimation sur petits domaines est proche de l'estimation directe. Lorsque l'estimation directe n'est pas fiable, ψ_i tend à être grande et l'estimation sur petits domaines est proche de l'estimation synthétique.

En pratique, la variance lissée $\tilde{\psi}_i$ n'est jamais connue et doit être remplacée par une estimation appropriée, $\hat{\tilde{\psi}}_i$. Pour les régions présentant une grande taille d'échantillon, l'estimation directe $\hat{\psi}_i$ de ψ_i obtenue à l'aide des méthodes standard, telles que le bootstrap, est généralement une estimation précise de la variance lissée $\tilde{\psi}_i$. C'est ce qui a été utilisé dans l'EPA pour les grandes régions. Cependant, pour les régions présentant une taille d'échantillon réduite, l'estimation directe $\hat{\psi}_i$ n'est généralement pas fiable. Pour résoudre ce problème, on a modélisé l'estimation de la variance $\hat{\psi}_i$ et sa valeur prédite a été utilisée comme estimation de la variance lissée. Cela a permis de stabiliser l'estimation de $\hat{\psi}_i$ aux dépens de l'introduction d'un autre modèle. De plus amples détails sur l'estimation de la variance lissée sont fournis dans Beaumont et Bocci (2016).

Trois intrants doivent être fournis au système EPD pour chaque région, afin d'obtenir des estimations sur petits domaines $\hat{\theta}_i^{EPD}$, $i = 1,...,M$:

i. des estimations directes $\hat{\theta}_i$;

ii. des estimations de la variance lissée $\hat{\tilde{\psi}}_i$;

iii. un vecteur de variables auxiliaires $\mathbf{z}_i$.

Nous décrivons ci-dessous comment le vecteur $\mathbf{z}_i$ a été défini pour l'estimation du taux de chômage et du nombre de personnes avec emploi.

Supposons que N_i^{emp} et N_i^{cho} soient respectivement le nombre de personnes employées et au chômage dans la région i. Le taux de chômage dans la région i est défini comme $\theta_i = \frac{N_i^{cho}}{N_i^{cho} + N_i^{emp}}$ alors que le nombre d'emplois dans la région i est $\theta_i = N_i^{emp}$. Trois variables auxiliaires sont disponibles : N_i^{bae}, le nombre de bénéficiaires de l'assurance-emploi dans la région i, N_i^{15-64} le nombre de personnes âgées entre 15 et 64 ans dans la région i, et N_i^{65+} le nombre de personnes âgées de 65 ans et plus dans la région i. Définissons le taux de bénéficiaires comme étant $Q_i = \frac{N_i^{bae}}{N_i^{15+}}$ et le taux de 15 à 64 comme $R_i = \frac{N_i^{15-64}}{N_i^{15+}}$. Pour l'estimation du taux de chômage, nous utilisons $\mathbf{z}_i = (1, Q_i, R_i)'$. La première composante du vecteur représente l'ordonnée à l'origine. Nous avons également utilisé des termes pour ajuster une spline linéaire dans Q_i et R_i. Pour l'estimation du nombre de personnes avec emploi, les considérations de modélisation nous ont incités à procéder différemment. Au lieu de modéliser N_i^{emp} directement, nous avons modélisé la proportion de personnes employées, $\frac{N_i^{emp}}{N_i^{15+}}$, à l'aide de la constante associée à l'ordonnée à l'origine, de Q_i et de R_i dans $\mathbf{z}_i$ comme pour le modèle du taux de chômage, mais nous avons également utilisé des termes pour ajuster une spline quadratique dans Q_i et R_i. L'estimation sur petits domaines de $\frac{N_i^{emp}}{N_i^{15+}}$ est alors multipliée par N_i^{15+} afin d'obtenir l'estimation sur petits domaines du nombre d'emplois N_i^{emp}.

Enfin, un processus de réconciliation a été appliqué aux estimations par petites régions du nombre de personnes avec emploi N_i^{emp} afin de s'assurer que, lorsqu'il est agrégé au niveau provincial, le total correspond à l'estimation provinciale directe du nombre de personnes employées. Il convient de noter que les régions ne s'agrègent pas exactement à la province, car certaines régions chevauchent deux provinces. Dans ces cas, la combinaison de régions la plus proche de la définition géographique de la province a été utilisée pour la réconciliation.

3. Évaluation des estimations sur petits domaines

L'exactitude des estimations sur petits domaines dépend de la fiabilité du modèle de Fay-Herriot. Il est donc essentiel d'évaluer attentivement la validité du modèle avant de diffuser toute estimation. Par exemple, il est important de vérifier qu'il existe une relation linéaire entre $\hat{\theta}_i$ et $\mathbf{z}_i$ au moins approximativement. Une manière simple de vérifier l'hypothèse de linéarité est de représenter graphiquement les résidus normalisés :

$$\hat{a}_i = \frac{\hat{\theta}_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}_v^2 + \hat{\tilde{\psi}}_i}}$$

en fonction des valeurs prédites $\mathbf{z}_i'\hat{\boldsymbol{\beta}}$. L'hypothèse linéaire est raisonnable lorsque le graphique ne révèle pas de tendance particulière. Les résidus normalisés sont des statistiques clés qui peuvent également être utilisées pour vérifier les autres hypothèses du modèle, comme la normalité des erreurs du modèle. À cet égard, nous avons élaboré un test fondé sur les résidus normalisés, afin de déterminer les régions aberrantes, c.-à-d., les régions qui ne suivent pas le même modèle que les autres régions. Les régions aberrantes entraîneraient généralement des résidus extrêmes. Nos évaluations de modèles pour l'EPA n'ont révélé aucune erreur évidente de spécification de modèle ni aucune région aberrante. Cela est rassurant, car la qualité des estimations sur petits domaines dépend considérablement de la fiabilité du modèle de Fay-Herriot.

L'erreur quadratique moyenne (EQM) est le concept habituel utilisé pour évaluer les gains d'efficacité provenant de l'utilisation de l'estimation sur petits domaines $\hat{\theta}_i^{EPD}$ plutôt que l'estimation directe $\hat{\theta}_i$:

$$\text{EQM}(\hat{\theta}_i^{EPD}) = E_{mp}(\hat{\theta}_i^{EPD} - \theta_i)^2 .$$

L'EQM est inconnue, mais peut être estimée. Lorsqu'aucune réconciliation n'est effectuée sur les estimations, on peut utiliser les formules données dans Rao et Molina (2015). Pour les estimations réconciliées finales des chiffres de l'emploi, les formules d'EQM appropriées ont été dérivées et utilisées (Verret et Walker, 2025). On s'attend à des gains d'efficacité sur l'estimation directe lorsque l'estimation de l'EQM est inférieure à l'estimation de la variance lissée $\hat{\tilde{\psi}}_i$ ou à l'estimation de la variance directe $\hat{\psi}_i$. En général, les estimations sur petits domaines de l'EPA étaient nettement plus efficaces que les estimations directes, en particulier pour les régions dont la taille de l'échantillon était la plus petite.

Nous avons également comparé les estimations directes et sur petits domaines pour mai 2016 avec les estimations de la population active du Recensement 2016, qui sont basées sur un échantillon de taille beaucoup plus importante. En moyenne, les estimations sur petits domaines étaient considérablement plus proches des estimations du recensement que des estimations directes (Beaumont et Bocci, 2016).

Afin de garantir la robustesse des modèles dans le temps, les modèles d'estimation sur petits domaines ont été élaborés et évalués en utilisant plus de deux ans de données mensuelles avant leur diffusion.

Références

Beaumont, J.-F. Et C. Bocci. (2016). Estimation sur petits domaines dans l'Enquête sur la population active. Document présenté au Comité consultatif sur les méthodes statistiques de Statistique Canada le 31 mars 2016.

Estevao, V., Y. You, M. Hidiroglou, J.-F. Beaumont (2023a). Estimations pour petits domaines – Modèle au niveau du domaine avec estimation EBLUP – Description des paramètres de fonction et guide de l'utilisateur. Document de Statistique Canada.

Estevao, V., Y. You, M. Hidiroglou, J.-F. Beaumont et S. Rubin-Bleuer. (2023b). Estimations pour petits domaines – Modèle au niveau du domaine avec estimation EBLUP – Spécifications méthodologiques. Document de Statistique Canada.

OECD (2020)., Delineating Functional Areas in All Territories, OECD Territorial Reviews, OECD Publishing, Paris, https://doi.org/10.1787/07970966-en.

Rao, J.N.K. et I. Molina. (2015). *Small Area Estimation*. John Wiley & Sons, Inc., Hoboken (New Jersey).

Verret, F. et Walker, B. (2025). Rétro-ingénierie d'un processus de réconciliation hypothétique pour estimer l'erreur quadratique moyenne des estimations sur petits domaines qui ont été réconciliées. *Recueil du Symposium international sur les questions de méthodologie*, Statistique Canada.