Nº 11-522-x au catalogue
ISSN : 1709822-X

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Création d'un indicateur composite de la qualité pour les estimations fondées sur des données administratives au moyen du partitionnement

par Roxanne Gagnon, Martin Beaulieu, Danielle Lebrasseur,
Wei Qian et Anthony Yeung

Date de diffusion : le 22 octobre 2021

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Création d'un indicateur composite de la qualité pour les estimations fondées sur des données administratives au moyen du partitionnement

Roxanne Gagnon, Martin Beaulieu, Danielle Lebrasseur, Wei Qian et Anthony Yeung[1]

Résumé

Les agences nationales de statistique telles que Statistique Canada se doivent de communiquer la qualité de l'information statistique aux utilisateurs. Les méthodes traditionnellement utilisées pour le faire sont fondées sur des mesures de l'erreur d'échantillonnage. Elles ne sont donc pas adaptées aux estimations produites à partir des données administratives pour lesquelles les sources d'erreur principales sont non dues à l'échantillonnage. Une approche plus adaptée à ce contexte pour rapporter la qualité des estimations présentées dans un tableau multidimensionnel est décrite dans cet article. Des indicateurs de qualité ont été dérivés pour diverses étapes de traitement post-acquisition, comme le couplage, le géocodage et l'imputation, par domaine d'estimation. Un algorithme de partitionnement a ensuite servi à regrouper les domaines présentant des niveaux de qualité similaires pour une estimation donnée. Des cotes visant à informer les utilisateurs sur la qualité relative des estimations d'un domaine à l'autre ont été attribuées aux groupes ainsi formés. Cet indicateur, nommé l'indicateur composite de la qualité (ICQ), a été développé et appliqué de façon expérimentale dans le cadre du Programme de la statistique du logement canadien (PSLC) qui a comme objectif la production de statistiques officielles sur le secteur du logement résidentiel au Canada par l'intégration de multiples sources de données administratives.

Mots Clés : Apprentissage automatique non supervisé, assurance de la qualité, données administratives, intégration des données, partitionnement.

1. Introduction

Les utilisateurs doivent être informés de la qualité des données pour leur permettre de juger si l'information statistique convient à l'usage qu'ils veulent en faire. Il s'agit d'un principe directeur de l'assurance qualité à Statistique Canada, qui est énoncé dans le *Cadre d'assurance de la qualité* (Statistique Canada, 2017) et dans la *Politique visant à informer les utilisateurs sur la qualité des données et la méthodologie* (Statistique Canada, 2000). Cette exigence est satisfaite de plusieurs manières, par exemple la diffusion des rapports techniques et des métadonnées relatives aux différentes enquêtes qui aident l'utilisateur à juger de la qualité dans toutes ses dimensions (pertinence, actualité, exactitude, cohérence, intelligibilité et accessibilité). Il est aussi recommandé de rapporter une mesure de l'exactitude de chaque estimation. Pour les enquêtes par échantillon, une mesure de l'erreur d'échantillonnage est diffusée, comme un intervalle de confiance, un coefficient de variation ou une cote allant de *A* à *F* dérivée du coefficient de variation.

Dans les dernières années, les agences nationales de statistique ont commencé la transition vers des modèles intégrés où les enquêtes probabilistes ne sont plus utilisées seules, mais en combinaison avec d'autres sources de données administratives ou alternatives (p. ex. mégadonnées, télédétection, moissonnage du web). Ces nouvelles sources de données peuvent être utilisées en complément des enquêtes par échantillon, être complétées par des enquêtes par échantillon ou même remplacer complètement les enquêtes par échantillon. L'intégration des données constitue une occasion exceptionnelle d'améliorer la qualité des statistiques officielles sur le plan de l'actualité et potentiellement de l'exactitude. Mesurer et rapporter l'exactitude représente toutefois un défi important dans ce contexte, car les méthodes et la terminologie utilisées par les agences nationales de statistique sont encore largement ancrées dans la théorie de l'échantillonnage.

[1]Roxanne Gagnon, Statistique Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6 (roxanne.gagnon@statcan.gc.ca) ; Martin Beaulieu, Statistique Canada (martin-j.beaulieu@statcan.gc.ca) ; Danielle Lebrasseur, Statistique Canada, (danielle.lebrasseur@statcan.gc.ca) ; Wei Qian, Statistique Canada, (wei.qian@statcan.gc.ca); Anthony Yeung, Statistique Canada, (anthony.yeung@statcan.gc.ca).

Des travaux ont été réalisés au sein d'autres organisations en vue d'adapter le paradigme de l'erreur totale d'enquête aux données intégrées (Zhang, 2012), aux données administratives (Reid et coll., 2017) et aux mégadonnées (Amaya et coll., 2020). Ils ont permis de recenser de façon exhaustive les sources d'erreur possibles et d'adapter la terminologie existante à ces nouvelles sources de données. Ils proposent toutefois peu de recommandations concrètes sur la façon de rapporter ces erreurs quantitativement et en fonction de leur impact sur les estimations afin d'informer les utilisateurs. À Statistique Canada, les bonnes pratiques recommandées dans les *Lignes directrices concernant la qualité* (Statistique Canada, 2019) consistent à dériver des indicateurs reflétant la qualité des processus de production d'information statistique pour juger de leur performance et de leur impact potentiel sur les estimations. Pris individuellement, ces indicateurs sont simples à comprendre et à mesurer. Lorsqu'ils sont considérés dans leur ensemble, il peut être difficile d'en tirer un portrait suffisamment clair pour pouvoir être communiqué aux utilisateurs. L'approche de l'indicateur composite de qualité (ICQ) propose d'utiliser l'apprentissage automatique non supervisé pour combiner les indicateurs en une valeur unique, analogue à celle utilisée couramment dans les enquêtes par échantillon.

2. Programme de la statistique du logement canadien

L'ICQ a été développé pour rapporter l'exactitude des estimations du Programme de la statistique du logement canadien (PSLC). Il s'agit d'un programme statistique ayant exclusivement recours à des sources de données administratives pour produire des statistiques visant à suivre et analyser le marché canadien du logement résidentiel. Ces données présentent des niveaux de qualité variables au moment de leur acquisition et les diverses étapes de traitement et de production des estimations finales peuvent éventuellement introduire des erreurs. Trois tableaux de résultats (Statistique Canada, 2021) comprenant plusieurs estimations ont été ciblés pour développer l'ICQ. Les domaines d'estimation sont définis en fonction de plusieurs variables catégoriques et d'une variable géographique présentée selon plusieurs niveaux, le plus détaillé étant la subdivision de recensement (SDR)[2].

3. Création de l'indicateur composite de la qualité

3.1 Sélection des indicateurs de qualité

La première étape a consisté à identifier les indicateurs de qualité des processus qui ont permis d'obtenir les variables impliquées dans la construction des différents tableaux. Dans l'exemple du PSLC, l'exactitude des variables qualitatives a été représentée par le taux de géocodage, le score de confiance moyen du géocodage, des taux de codage et des taux d'erreur de couplage d'enregistrements, c'est-à-dire le taux de fausses découvertes (TFD) et le taux de faux négatifs (TFN). L'exactitude des variables quantitatives a été représentée par des taux de déclaration et d'inclusion. Ces indicateurs sont résumés dans le tableau 3.1-1. Ils sont estimés au niveau des domaines.

Tableau 3.1-1
Indicateurs de qualité sélectionnés

Variables[3]	Indicateur de qualité
Géographie	Taux de géocodage, Score de confiance moyen du géocodage
Période de construction	Taux de codage
Superficie habitable totale	Taux de déclaration, Taux d'inclusion
Type de propriétaire	Taux de codage
Type de propriété	Taux de codage
Usage de la propriété	Précision (1 - TFD), Rappel (1 - TFN)
Valeur de l'évaluation foncière	Taux de déclaration

[2] *Dictionnaire du Recensement de 2016*, https://www12.statcan.gc.ca/census-recensement/2016/ref/dict/geo012-fra.cfm
[3] *Base de métadonnées intégrées*, https://www23.statcan.gc.ca/imdb/p2SV_f.pl?Function=getSurvVariableList&Id=1322247

3.2 Partitionnement

Les indicateurs de qualité ont été combinés en utilisant le partitionnement à k-moyennes. Cet algorithme a permis de grouper les domaines d'estimation qui présentent des niveaux semblables pour les différents indicateurs. Il exploite une mesure de distance qui représente la dissimilarité entre les domaines d'estimation, en considérant tous les indicateurs de qualité simultanément. Un graphique diagnostique a été utilisé pour choisir le nombre optimal de partitions. Les autres spécifications du modèle, c'est-à-dire le choix des indicateurs de qualité à inclure ainsi que les poids assignés à chaque indicateur, ont été déterminées en fonction du type d'estimation et de la variable d'intérêt.

3.3 Pondération

Les indicateurs de qualité ont d'abord été standardisés pour retirer l'effet de l'échelle, puis pondérés dans le but de contrôler l'importance accordée à chaque indicateur lors du partitionnement. Deux méthodes de pondération ont été utilisées, une pour les comptes et les pourcentages et l'autre pour les variables d'intérêt continues.

Dans le cas des comptes et des pourcentages, des poids égaux ont été assignés à chaque variable de domaine, puis le poids de chaque variable a été réparti également entre les indicateurs de qualité reliés à cette variable. Soit un tableau construit à partir de D variables de domaine désignées comme X_d pour d = 1, …, D. L'exactitude de chaque variable X_d est représentée par les indicateurs de qualité $IQ_{X_{dj}}$ pour j = 1, …, J_{X_d}. Le poids $w_{X_{dj}}$ de $IQ_{X_{dj}}$ est donné par :

$$w_{X_{dj}} = \frac{1}{D \cdot J_{X_d}}$$

Dans le cas des totaux, des moyennes et des médianes des variables d'intérêt continues, les poids des variables de domaine ont été répartis proportionnellement selon la force de la relation entre la variable de domaine et la variable d'intérêt. Soit un tableau de résultats construit à partir d'une variable d'intérêt continue Y et de D variables de domaine. Le poids des indicateurs de qualité IQ_{Y_j} pour j = 1, …, J_Y est donné par w_{Y_j} :

$$w_{Y_j} = \frac{1}{(D+1) \cdot J_Y}$$

Les poids des indicateurs de qualité $IQ_{X_{dj}}$ est donné par :

$$w_{X_{dj}} = \frac{D}{(D+1)} \cdot \frac{\eta_{X_d}}{\sum_{d=1}^{D} \eta_{X_d}} \cdot \frac{1}{J_{X_d}}$$

où η_{X_d} correspond à la taille de l'effet de la variable X_d dans le modèle d'ANOVA de Y en fonction des variables de domaine. Le tableau 3.3-1 présente les résultats des modèles d'ANOVA réalisés à partir des microdonnées (propriétés) pour les variables de domaine et d'intérêt des tableaux sélectionnés du PSLC.

Tableau 3.3-1
Résultats des modèles d'ANOVA

Effet (X_d)	Taille de l'effet sur la variable d'intérêt Y (%)		
	Valeur de l'évaluation foncière	Surface habitable totale	Valeur de l'évaluation foncière au pied carré
Période de construction	1,1	5,7	0,9
SDR	26,1	9,9	26,7
Type de propriétaire	0,2	0,2	0,0
Type de propriété	7,2	24,4	0,2
Usage de la propriété	0,0	0,4	0,1
Résiduel	65,3	59,4	72,0

Comme on peut le voir au tableau 3.3-1, la SDR était la variable de domaine qui expliquait la plus grande partie de la variance (26,1 %) observée de la valeur de l'évaluation foncière. Ce sont donc les indicateurs de qualité reliés à la SDR qui ont été les plus importants lors de la création de l'ICQ pour les estimations de la moyenne et la médiane de la valeur de l'évaluation foncière. Dans le cas de la moyenne et la médiane de la surface habitable totale, ce sont plutôt les indicateurs de qualité reliés au type de propriété qui devaient être les plus importants puisque cette variable expliquait la plus grande partie de la variance (24,4 %) de la surface habitable totale. L'ensemble des variables de domaine n'expliquaient qu'une partie de la variance des trois variables d'intérêt. La taille des effets pourrait donc être biaisée par l'omission de certaines variables. Cependant, vu que l'objectif était d'obtenir une mesure de l'importance relative des variables de domaine dans l'estimation et non de réaliser un modèle explicatif ou prédictif, il n'était pas utile d'ajouter d'autres variables ou interactions au modèle.

3.4 Ordre des groupes

Les groupes obtenus par partitionnement n'étaient pas ordonnés. Un ordre leur a donc été assigné pour faciliter l'interprétation. Cet ordre a été établi en fonction d'un score global calculé pour chaque groupe k de la façon suivante :

$$\text{Score}_k = \frac{\sum_{i=1}^{M} \mathbb{I}_{ik} \cdot \overline{\text{IQ}}_i}{\sum_{i=1}^{M} \mathbb{I}_{ik}}$$

où M est le nombre total de domaines d'estimation, $\mathbb{I}_{ik}$ est une fonction qui prend la valeur 1 si le domaine d'estimation i a été assigné au groupe k, 0 sinon, et $\overline{\text{IQ}}_i$ est la moyenne pondérée des indicateurs de qualité :

$$\overline{\text{IQ}}_i = \sum_{\text{j}=1}^{J_Y} w_{Y_j} \cdot IQ_{Y_{ji}} + \sum_{d=1}^{D} \sum_{\text{j}=1}^{J_{X_d}} w_{X_{dj}} \cdot IQ_{X_{dji}}$$

Le groupe avec le plus haut score global a été désigné comme le groupe A, le groupe avec le second plus haut score est désigné comme le groupe B, etc.

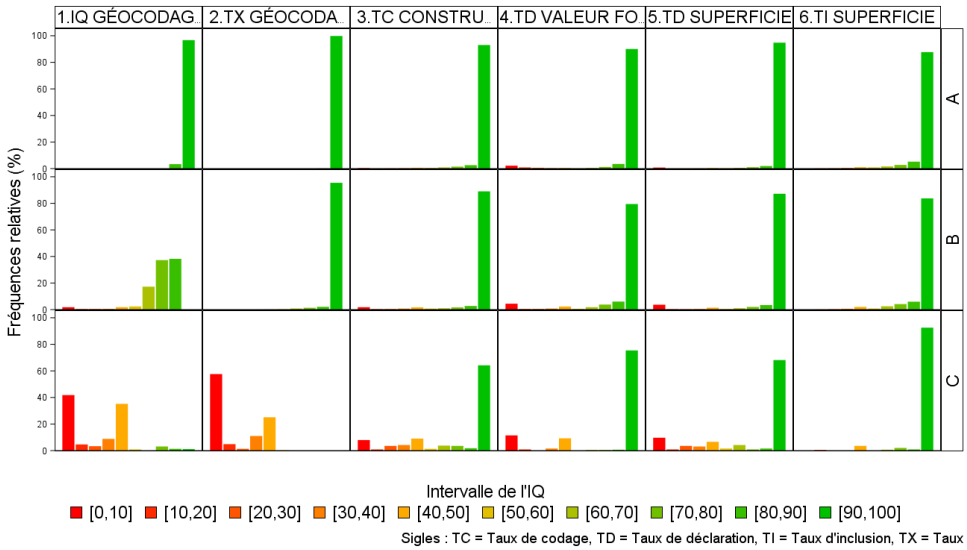
3.5 Visualisation des profils des groupes

Afin de pouvoir interpréter correctement les valeurs de l'ICQ, il est impératif de procéder à une visualisation des résultats pour comprendre ce qui distingue les groupes obtenus entre eux. La figure 3.5-1 à la page suivante est un exemple de visualisation développée pour les besoins de l'étude. Chaque cellule de la grille présente la distribution de fréquences relatives d'un indicateur de qualité donné, parmi les colonnes, et pour un groupe donné, parmi les lignes.

On voit à la figure 3.5-1 que les groupes A et B se distinguent entre eux par le score de confiance moyen du géocodage (première colonne). Il est presque toujours au-dessus de 90 % pour le groupe A (première ligne) et entre 60 % et 90 % pour le groupe B (seconde ligne). Le groupe C (troisième ligne) se distingue des deux autres groupes à la fois par le score de confiance moyen du géocodage (première colonne) et le taux de géocodage (deuxième colonne) qui sont tous les deux inférieurs à 50 %.

Dans le contexte du PSLC, plusieurs modèles de partitionnement ont été nécessaires pour décrire l'exactitude des estimations, car les tableaux de résultats sélectionnés incluaient plusieurs types d'estimation et plus d'une variable d'intérêt. Une visualisation des données comme celle présentée à la Figure 3.5-1 a été réalisée pour chaque modèle de partitionnement requis. Les cotes finales assignées à chaque groupe de chaque modèle ont été déterminées après comparaison des graphiques des différents modèles.

Figure 3.5-1
Exemple de visualisation des données — Fréquences relatives des indicateurs de qualité dans chaque groupe



3.6 Attribution des étiquettes finales et interprétation

Le Tableau 3.6-1 présente l'information fournie aux utilisateurs pour les aider à juger de l'adéquation à l'utilisation des estimations du PSLC. Chaque valeur de l'ICQ est associée à une étiquette standard. Une description des indicateurs de qualité, désignés comme les composantes de l'ICQ, est fournie pour chaque groupe.

Tableau 3.6-1
Valeurs de l'indicateur composite de qualité, étiquette et description des composantes

Valeur ICQ	Étiquette	Description des composantes
A	Excellent	Toutes les composantes de l'ICQ sont jugées excellentes.
B	Très bon	L'ICQ est jugé très bon en raison du très bon niveau de qualité du géocodage et de l'excellent niveau de qualité des autres composantes.
C	Bon	L'ICQ est jugé bon en raison du bon niveau de qualité du géocodage et de l'excellent niveau de qualité des autres composantes.
D	Acceptable	L'ICQ est jugé acceptable en raison du bas niveau de qualité du géocodage ou du codage de la période de construction tandis que le niveau de qualité des autres composantes est excellent.
E	Utiliser avec prudence	L'ICQ indique un niveau de qualité qui incite à la prudence lors de l'utilisation.
F	Trop peu fiable pour être publié	-

L'utilisation du partitionnement pour créer l'ICQ implique que les valeurs obtenues sont considérées comme relatives et non absolues, c'est-à-dire que l'exactitude d'une estimation dans un domaine n'est pas évaluée par rapport à un standard préétabli, mais plutôt en comparaison des estimations dans les autres domaines du tableau. Il serait difficile d'établir des standards pour différents programmes, tableaux, types d'estimation ou variables d'intérêt qui seraient à la fois cohérents et pertinents, puisque les indicateurs de qualité et les poids ne sont pas toujours les mêmes d'une estimation à l'autre. Malgré cela, le partitionnement est la manière la plus objective de créer des groupes de qualité et d'éviter de recourir à des règles arbitraires.

4. Limites

Les limites de l'ICQ concernent surtout les indicateurs de qualité utilisés pour le partitionnement. Dans le contexte du PSLC, certains indicateurs mesurent la proportion des unités administratives pour lesquelles une étape de traitement a été complétée, mais pas nécessairement la fiabilité avec laquelle l'étape a été réalisée. Par exemple, une valeur peut être codée, mais on ne connaît pas la probabilité que la valeur codée soit exacte. Pour ce qui est des taux d'erreur du couplage d'enregistrements, la mesure de ces taux nécessite la vérification manuelle d'un échantillon. Il serait long et dispendieux de procéder à la sélection et la vérification d'un échantillon suffisamment grand pour les obtenir à l'échelle des SDR. Dans le contexte du PSLC, les taux d'erreur de couplage n'étaient disponibles qu'à des niveaux géographiques supérieurs, ce qui a limité l'efficacité de ces indicateurs de qualité dans les modèles de partitionnement.

5. Conclusion

Le partitionnement est une approche simple, rapide et efficace pour combiner une série d'indicateurs de qualité basés sur les étapes de traitement des données comme le géocodage, le couplage d'enregistrements et l'imputation. Les indicateurs de qualité peuvent être pondérés avant le partitionnement dans le but d'accorder plus d'importance à certains d'entre eux. Dans l'exemple du PSLC, les poids ont été obtenus de façon objective grâce à la modélisation des variables d'intérêt continues. Les groupes obtenus rassemblent des domaines d'estimation qui sont affectés par des enjeux de qualité similaires pour une estimation donnée. Ceci permet d'offrir aux utilisateurs une description des composantes de qualité en plus de la valeur de l'ICQ et d'une étiquette pour rapporter l'exactitude de l'estimation. L'ICQ a été diffusé pour la première fois en septembre 2021 dans certains tableaux du PSLC. C'est la première fois qu'une cote représentant l'exactitude des estimations est proposée aux utilisateurs dans le cadre d'un programme statistique entièrement basé sur l'intégration des données administratives à Statistique Canada.

Bibliographie

Amaya, A., Biemer, P. et Kinyon, D. (2020), « Total Error in a Big Data World: Adapting the TSE Framework to Big Data », Journal of Survey Statistics and Methodology, 8 (1), pp. 89–119. https://doi.org/10.1093/jssam/smz056

Reid, G., Zabala, F. et Holmberg, A. (2017), « Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ », Journal of Official Statistics, 33 (2), pp. 477—511. http://dx.doi.org/10.1515/JOS-2017-0023

Statistique Canada (2000), *Politique visant à informer les utilisateurs sur la qualité des données et la méthodologie*, Ottawa, Ontario. https://www.statcan.gc.ca/fr/apercu/politique/info-usager

Statistique Canada (2017), *Le cadre d'assurance de la qualité de Statistique Canada*, Catalogue no. 12-586-X, Ottawa, Ontario. https://www150.statcan.gc.ca/n1/fr/catalogue/12-586-X

Statistique Canada (2019), *Statistique Canada : Lignes directrices concernant la qualité*, Sixième édition, Catalogue no. 12-539-X, Ottawa, Ontario. https://www150.statcan.gc.ca/n1/pub/12-539-x/12-539-x2019001-fra.htm

Statistique Canada (2021), *Tableau 46-10-0027-01 Participation à la résidence des propriétés résidentielles, par type de propriété et période de construction*. https://doi.org/10.25318/4610002701-fra

Statistique Canada (2021), *Tableau 46-10-0053-01 Type de propriétaire et usage de la propriété par type de propriété résidentielle et période de construction.* https://doi.org/10.25318/4610005301-fra

Statistique Canada (2021), *Tableau 46-10-0054-01 Résidence de la propriété et usage de la propriété par type de propriété résidentielle et période de construction.* https://doi.org/10.25318/4610005401-fra

Zhang, L.-C. (2012), « Topics of Statistical Theory for Register-Based Statistics and Data Integration », Statistica Neerlandica, 66 (1), pp. 41–63. http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x