Nº 11-522-X au catalogue
ISSN 1709-822X

Recueil du Symposium de 2022 de Statistique Canada : Désagrégation des données : dresser un portrait de données plus représentatif de la société

Intégration des données existantes pour élaborer un indicateur d'ethnicité dans le cadre du PEDSL

par Aziz Farah, Bassirou Diagne et Abdelnasser Saïdi

Date de diffusion : le 25 mars 2024

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2022 de Statistique Canada
Désagrégation des données : Dresser un portrait de données plus représentatif de la société

Intégration des données existantes pour élaborer un indicateur d'ethnicité dans le cadre du PEDSL

Aziz Farah, Bassirou Diagne et Abdelnasser Saïdi[1]

Résumé

Le Programme d'élaboration de données sociales longitudinales (PEDSL) est une approche d'intégration des données sociales destinée à fournir des opportunités analytiques longitudinales sans imposer un fardeau de réponse supplémentaire aux répondants. Le PEDSL tire parti d'une multitude de signaux qui proviennent de différentes sources de données pour la même personne, ce qui permet de mieux comprendre leurs interactions et de suivre l'évolution dans le temps. Cet article traitera de la façon dont le statut d'ethnicité des personnes au Canada peut être estimé au niveau désagrégé le plus détaillé possible en utilisant les résultats d'une variété de règles opérationnelles appliquées aux données déjà appariées et au dénominateur du PEDSL puis montrera comment des améliorations ont pu être obtenues en utilisant des méthodes d'apprentissage automatique telles que des arbres de décision et des techniques de forêt aléatoire.

Mots Clés : Intégration des données; Apprentissage automatique; Indicateur sur l'ethnicité.

1. Introduction

Dans le contexte du plan d'action sur les données désagrégées (PADD), le Programme d'élaboration de données sociales longitudinales (PEDSL) propose de développer un algorithme qui a pour objectif d'élaborer un indicateur sur l'ethnicité pour chaque individu de la population canadienne. L'algorithme porte sur l'année de référence 2016 et peut être généralisé à toute année ultérieure. L'assignation sera faite au niveau individuel (le niveau le plus désagrégé possible), pour tous les individus de la population canadienne. Cet indicateur fournira des opportunités analytiques additionnelles lorsqu'utilisé comme covariable. De plus il permettra de développer des bases de sondage plus inclusives lorsqu'on veut cibler des groupes ethniques particuliers pendant les opérations de répartition ou de stratification de l'échantillon.

Le PEDSL intègre des données administratives déjà existantes et appariées selon une clé unique et anonyme construite dans l'Environnement de couplage des données sociales (ECDS). L'ECDS est un environnement sécurisé respectant en permanence les normes les plus strictes en matière de la vie privée et de la sécurité des données. De même le PEDSL est un ensemble d'algorithmes et de processus destinés à l'analyse des trajectoires de données sociales. Il n'est pas une vaste base de données intégrées, ni un environnement évolutif et ne contient aucun identificateur personnel.

2. Contexte

2.1 Catégories ethniques

Le concept d'ethnicité mesuré ici est similaire à celui de "l'identité ethnique" telle que définie dans la question sur le groupe de la population du Recensement de la population canadienne 2016 (question 19 du recensement) (veuillez vous référer à Statistique Canada (2016)). En plus des 12 groupes appartenant à des minorités visibles, deux autres catégories ne faisant pas partie des minorités visibles, le groupe "Blancs" et le groupe "Autochtones" y sont ajoutées pour former les 14 catégories ethniques suivantes:

[1]Aziz Farah, Statistique Canada, Canada, aziz.farah@statcan.gc.ca; Bassirou Daigne, Statistique Canada, Canada, bassirou.diagne@statcan.gc.ca ; Abdelnasser Saïdi, Statistique Canada, Canada, abdelnasser.saidi@statcan.gc.ca.

Table 2.1-1
Groupes ethniques

Code	Désignation française	
1	Sud-Asiatiques	Minorité visible
2	Chinois	
3	Noirs	
4	Philippins	
5	Latino-Américains	
6	Arabes	
7	Asiatiques du Sud-Est	
8	Asiatiques occidentaux	
9	Coréens	
10	Japonais	
11	Minorités visibles, n.i.a. (non inclus ailleurs)	
12	Minorités visibles multiples	
13	Autres, non-membres d'une minorité visible [Blancs]	Non-membres d'une minorité visible
14	Autochtones	

2.2 Dénominateur du PEDSL (Population canadienne)

C'est un portrait de la population canadienne d'intérêt pour une date de référence donnée. Il est produit sur demande au moyen d'un algorithme qui combine plusieurs sources de données déjà appariées selon l'ECDS. Ce dénominateur est au cœur des programmes de désagrégation et d'élaboration des données du PEDSL (veuillez vous référer à Aubin, P. (2021)). Le but ultime de ce travail est d'assigner à chaque individu du dénominateur du PEDSL l'une des 14 catégories ethniques de la Table 2.1-1 tout en maintenant des hauts niveaux de qualité.

2.3 Sources de données et signaux directs et indirects sur l'ethnicité

Les signaux directs (par autoréponse) ou indirects (par déduction) provenant des sources de données suivantes ont été utilisés pour le développement de l'indicateur:

- Les données des cinq derniers recensements qui contiennent les réponses au questionnaire détaillé (long) à la question sur le groupe de la population, à savoir les Recensements 2016, 2011, 2006, 2001 et 1996 (signal direct)
- Le fichier parent centrique (pour les années1993-2017) qui décrit la correspondance entre les parents et leurs enfants (signal indirect). Ce fichier est construit dans l'environnement ECDS (veuillez vous référer à Cascagnette, P. (2020, version 35), Cascagnette, P. (2020, version 30), Labrecque-Synnott, F. (2020), Gissler, G. (2020)).
- Le fichier historique des naissances et le fichier historique d'immigration qui représentent la source principale du dénominateur PEDSL (signal indirect).
- Le système d'information sur les apprentis inscrits (SIAI, 2008 et plus) (signal direct)
- Le système d'information sur les étudiants postsecondaires (SIEP, 2009 et plus) (signal direct)
- Le système d'information Ontarien sur la santé mentale (SIOSM, 2005 et plus) (signal direct)
- Le fichier des familles T1 (FFT1, 2005 et plus) (signal indirect)

3. Méthodologie

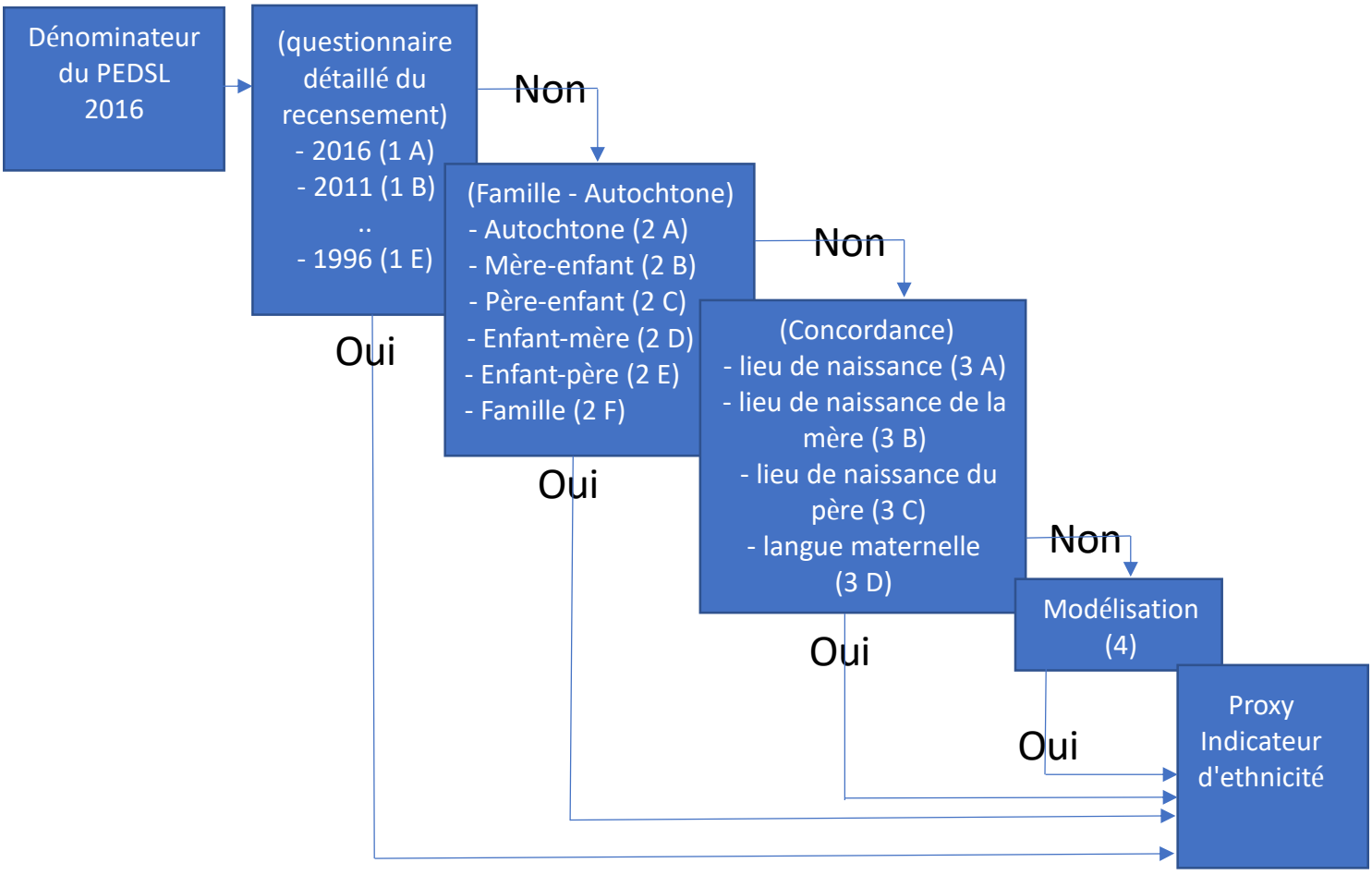
3.1 Méthodes directes et indirectes

Toutes les sources de données utilisées sont déjà appariées dans l'environnement ECDS. Cela permet d'associer au même individu plusieurs signaux relatifs à l'ethnicité (signaux directs ou indirects). Cette association peut être longitudinale (selon plusieurs années) et/ou selon les différentes sources de données. De plus cette association n'utilise aucun identifiant personnel mais plutôt une clé commune et anonyme issue de l'ECDS, ce qui rend la procédure complétement confidentielle.

Le même individu peut se retrouver dans les réponses de plusieurs recensements en même temps, avec des statuts ethniques identiques ou parfois conflictuels. Le même individu peut se retrouver également dans un fichier administratif (par exemple fichier historique d'immigration) envoyant un signal indirect qui permet de déduire son statut ethnique. Par exemple, le lieu de naissance ou la langue maternelle peuvent être considérés comme des signaux indirects lorsqu'ils réfèrent avec beaucoup d'évidence à une ethnicité en particulier.

Le Graphique 3.1-1 suivant montre la hiérarchie des étapes utilisées pour l'élaboration de l'indicateur (étape 1 puis étape 2, ...étape 4). À l'intérieur de chaque étape, les résultats de la classification de la sous-étape A sont retenus d'abord, puis la sous-étape B, puis C, …

Graphique 3.1-1
Hiérarchie des étapes de classification de l'indicateur sur l'ethnicité



La méthodologie utilisée assume que le recensement est la source d'information la plus fiable et par conséquent l'autoréponse provenant du recensement détaillé sera priorisée parmi tous les autres signaux directs ou indirects.

La Table 3.1-1 suivante, issue d'une étude de fluidité entre les réponses aux recensements 2016 et 2011, illustre la grande stabilité des réponses pour les principaux groupes ethniques d'un cycle de recensement à un autre (veuillez vous référer à Farah, A. (2022)).

Table 3.1-1
Taux de stabilité de réponse entre le recensement 2011 et 2016

Groupe	Taux de Stabilité	Groupe	Taux de Stabilité
Sud-Asiatiques	94,1%	Asiatiques du Sud-Est	73,5%
Chinois	95,8%	Asiatiques occidentaux	78,8%
Noirs	91,3%	Coréens	98,0%
Philippins	94,7%	Japonais	91,3%
Latino-Américains	84,9%	Blancs	97,7%
Arabes	80,5%	Autochtones	95,9%

La première étape du processus d'élaboration consiste à utiliser le maximum possible d'information directe provenant des recensements disponibles. Dans ce cas, lorsque le même individu répond à plusieurs recensements en même temps, la réponse au recensement le plus récent est retenue.

Les cas non classifiés par aucun des 5 derniers recensements disponibles, passeront à l'étape 2.
Dans cette étape, en assumant que l'ethnicité est relativement statique à l'intérieur d'une même famille, on peut procéder par l'imputation du statut ethnique d'un membre de famille, lorsque disponible, à un autre membre de la famille non encore classifié (parent, enfant, frère, sœur). Le fichier parent centrique permet de décrire la correspondance entre parents et enfants et par conséquent d'identifier l'appartenance à la même famille. L'approche d'imputation utilisée peut être critiquable et c'est pour cette raison que l'imputation familiale a été hiérarchisée à la deuxième étape.
Par la suite, les cas non classifiés passeront à l'étape 3 où on propose d'utiliser la correspondance entre les lieux de naissance, la langue maternelle et le statut ethnique pour quelques groupes en particulier. Dans ce cas, des règles de décision très strictes ont été développées pour ne retenir que quelques lieux de naissance (pays de naissance de l'individu ou de ses parents) et quelques langues maternelles dont le taux de bonne classification avec une ethnicité donnée dépasse 80 %. Plus concrètement les réponses pour le statut ethnique du Recensement de 2016 ont été croisées avec la classification issue de l'étape 3 pour sélectionner des sous-ensembles de pays ou de langues maternelles dont plus de 80 % de personnes ont répondu appartenir à l'ethnicité du Recensement 2016.

3.2 Modélisation et méthodes d'apprentissage automatique

Après l'application des méthodes directes et indirectes, il reste environ 30 % du dénominateur PEDSL non encore classifié. Par comparaison à la distribution du statut ethnique dans le recensement détaillé de 2016, on en déduit que la plupart de ces cas font partie du groupe "Blancs", du groupe "Noirs" et du groupe "Autochtones".

Le recours à la modélisation par apprentissage automatique s'avère utile après l'étape 3.
Les modèles utilisés ont été développés sur des données d'entrainement à hauteur de 70% en vue d'en optimiser les paramètres alors que les éléments de mesure de la qualité de prédiction ont été calculés avec des données de test.

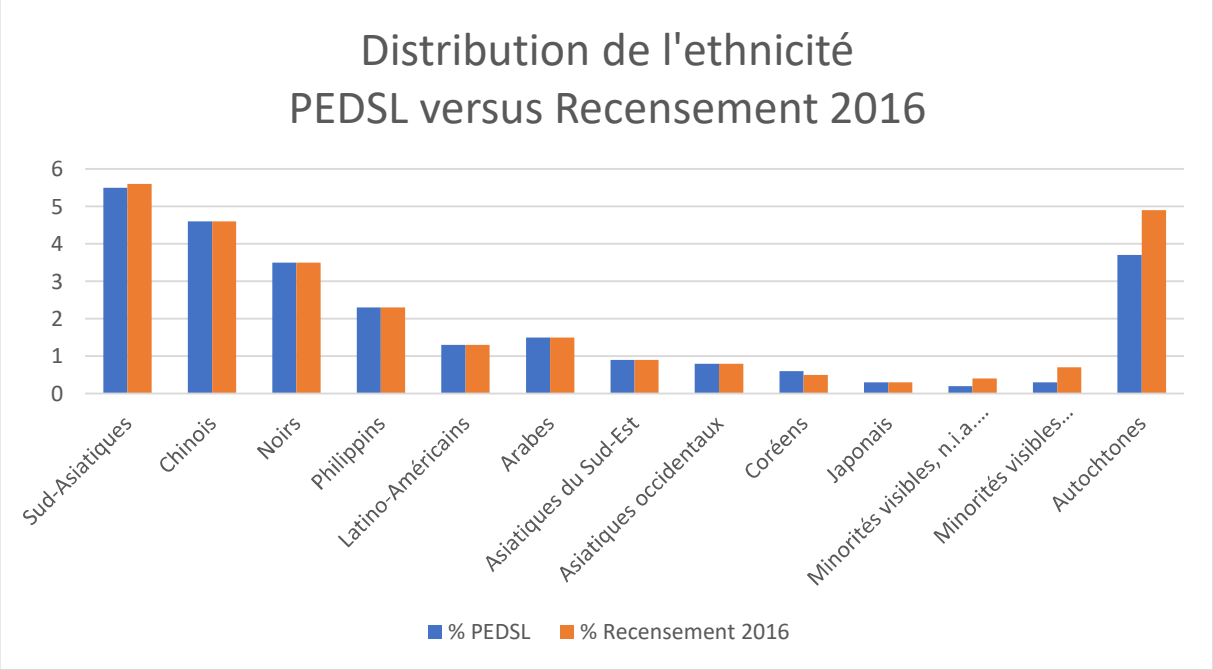
Puisque la majorité des cas non classifiés appartiennent à la catégorie "Blancs", le premier modèle de classification a été dédié à la classification exclusive de cette catégorie en utilisant la méthode de l'arbre de décision (veuillez vous référer à Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984)) avec l'option HPSPLIT dans SAS EG (veuillez vous référer à SAS Institute Inc. (2020)). Le deuxième modèle a été appliqué à la caractérisation des autres catégories ("Noirs", "Philippins", "Asiatiques du Sud Est" et "Japonais") à l'aide de l'algorithme XgBoost ou "boosting" de gradient extrême (veuillez vous référer à Chen, T., & Guestrin, C. (2016)). Ce modèle d'apprentissage permet un traitement rapide et efficace des données volumineuses et complexes par le biais d'une combinaison d'arbres de décision.

4. Résultats empiriques

4.1 Comparaison avec le Recensement 2016

À l'issue de toutes ces étapes, la comparaison de la distribution des comptes d'ethnicité entre le PEDSL et le Recensement 2016 montre une grande similitude pour la plupart des catégories à l'exception du groupe "Autochtones" auquel on n'a pas appliqué les méthodes de modélisation en plus des deux groupes difficiles "Minorités visibles multiples" et "Minorités visibles non inclus ailleurs". Le Graphique 4.1-1 suivant montre les résultats de cette comparaison. À noter que le groupe "Blancs" qui représente la catégorie la plus grande parmi toutes les autres catégories n'a pas été représenté dans cette figure pour des fins de de visibilité.

Graphique 4.1-1
Comparaison de la distribution de l'ethnicité entre les estimations du PEDSL et les estimations du Recensement 2016

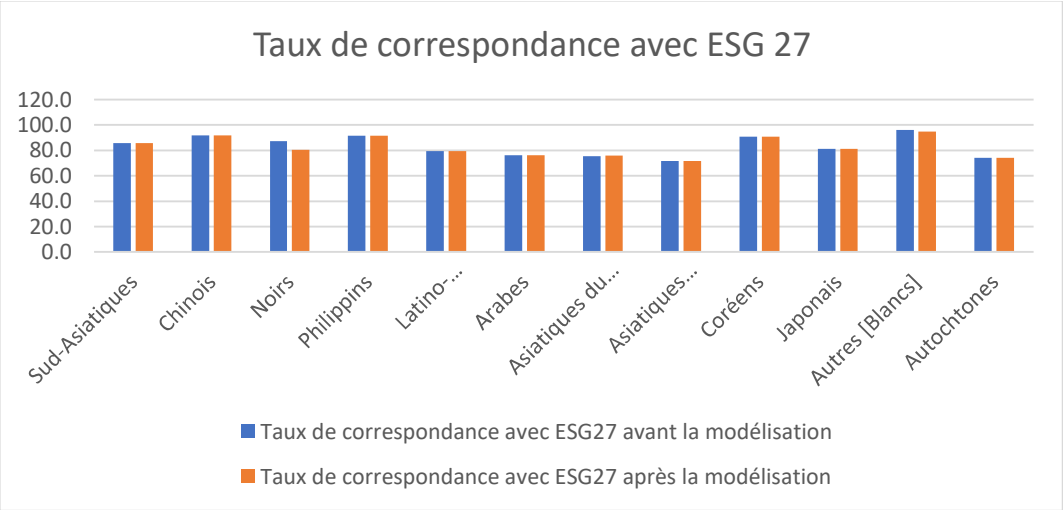


4.2 Micro-comparaison avec ESG 27 :

Le cycle 27 de l'Enquête sociale générale ESG a aussi porté sur l'ethnicité et ses données n'ont pas été utilisées pour le développement de l'indicateur sur l'ethnicité mais pour la validation des résultats de notre classification.

Le Graphique 4.2-1 suivant illustre les taux de correspondance (bonne classification) entre les estimations PEDSL et les estimations de l'ESG 27 avant et après l'utilisation de la modélisation.

Graphique 4.2-1
Validation des estimations avec ESG 27



5. Conclusion et perspectives d'avenir

L'utilisation des données appariées déjà existantes montre qu'il est possible de construire des indicateurs statiques comme l'ethnicité en utilisant des méthodes directes, indirectes et la modélisation par apprentissage automatique. Avec juste l'utilisation des données de recensements existants, environ 50 % de la population canadienne (Dénominateur PEDSL) a été classifié selon l'ethnicité. Les méthodes indirectes ont pu ajouter environ 20 % additionnel et la modélisation a permis d'atteindre presque la totalité de la population cible tout en maintenant des bons niveaux de qualité dans nos estimations.

Cependant il reste des défis à soulever pour les groupes les plus difficiles à classifier comme le groupe "Autochtones" et les deux groupes "Minorités visibles multiples" et "Minorités visibles non inclus ailleurs". Aussi, il serait pertinent d'approfondir l'étude sur l'origine de la surestimation et de la sous-estimation de certains groupes et de la possibilité d'appliquer des méthodes de calibration. La validation avec les données de recensements et avec une enquête externe (ESG) permet certes de justifier le niveau de la qualité des estimations en général, mais il faut aussi penser à développer un indicateur de qualité au niveau individuel.

6. Références

Aubin, P. (2021), « Methodology of the Longitudinal Social Data Development Program (LSDDP) Phase III », rapport non publié, Statistique Canada.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Cascagnette, P. (2020), " Linkage between Birth Fathers (1993 to 2017) and the SDLE Derived Record Depository", version 30, Document interne, Statistique Canada

Cascagnette, P. (2020), " Linkage between Birth Fathers (1993 to 2017) and the SDLE Derived Record Depository ", version 35, Document interne, Statistique Canada

Chen, T., & Guestrin, C. (2016), “XGBoost: A Scalable Tree Boosting System”, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 785–794, New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Farah, A. (2022), « Développement d'un indicateur sur les minorités visibles «, Document interne, Statistique Canada

Gissler, G. (2020), “ Linkage between Stillbirths_Mothers (1993 to 2017) and the SDLE Derived Record Depository ”, version 30, , Document interne, Statistique Canada

Labrecque-Synnott, F. (2020), “ Linkage between Stillbirths_Fathers (1993 to 2017) and the SDLE Derived Record Depository “, version 35, Document interne, Statistique Canada

SAS Institute Inc. (2020), SAS/STAT® 15.2, *User’s Guide*, Cary, NC: SAS Institute Inc.

Statistique Canada (2016), « Recensement de la population (2016) », Guide de référence sur les minorités visibles et le groupe de population (statcan.gc.ca).