

Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

Filer sur la voie du fastText : exploiter l'apprentissage automatique restreint par programmation linéaire pour réviser les classifications

par Justin Evans et Laura Wile

Date de diffusion : le 8 septembre 2025



Statistique
Canada

Statistics
Canada

Canada

Filer sur la voie du fastText : exploiter l'apprentissage automatique restreint par programmation linéaire pour réviser les classifications

Justin Evans et Laura Wile¹

Résumé

L'Enquête sur la population active (EPA) de Statistique Canada joue un rôle fondamental dans l'estimation des conditions du marché du travail au Canada. Périodiquement, l'EPA révisé ses données en fonction des versions les plus récentes des classifications des industries et des professions. Ces versions peuvent présenter des écarts importants, dont des changements structurels, des créations, des suppressions, des fractionnements et la combinaison d'unités de classification (catégories) au niveau général ainsi qu'au niveau du groupe de base. Par le passé, afin de rapprocher les catégories fractionnées, en vertu desquelles une catégorie était fractionnée en plusieurs catégories, on recodait manuellement un échantillon d'enregistrements fractionnés de l'EPA selon la nouvelle version de la classification. Selon la proportion de fractionnement observée dans l'échantillon recodé, une méthode d'allocation aléatoire était appliquée à toutes les données pour tenir compte de l'évolution du marché du travail canadien au fil du temps. Le présent article propose d'utiliser l'apprentissage automatique (fastText), contraint à des proportions fractionnées par programmation linéaire, pour réviser les classifications des industries et des professions dans l'EPA. Le cadre hybride bénéficie d'un mécanisme de révision basé sur le texte, tout en respectant les estimations traditionnelles des proportions, et en assurant une répercussion minimale sur la comparabilité des indicateurs du marché du travail publiés.

Mots clés : Apprentissage automatique; Enquête sur la population active; révision en fonction des classifications; fastText.

1. Introduction

L'Enquête sur la population active (EPA) de Statistique Canada joue un rôle fondamental dans l'estimation des conditions du marché du travail au Canada. De manière périodique, on révisé les données de l'EPA pour tenir compte des versions les plus récentes des classifications et de la désaisonnalisation, ou pour apporter des améliorations à la méthodologie. Grâce aux révisions, on s'assure que les estimations de l'enquête continuent de tenir compte de la taille et de la composition du marché du travail canadien, tout en minimisant les répercussions sur les tendances au chapitre des principaux indicateurs de marché, comme l'emploi, le chômage et les taux d'activité.

En janvier 2023, l'EPA est passée de la version 1.3 de la Classification nationale des professions (CNP) 2016 à la version 1.0 de la CNP 2021 pour le codage des professions. Ces versions présentaient des différences considérables, dont des changements structurels, des créations, des suppressions et la combinaison d'unités de classification (catégories) au niveau général ainsi qu'au niveau du groupe de base (Statistique Canada, 2023). Parmi les aspects les plus difficiles à rapprocher, il y avait les unités de classification fractionnées (tableau 1.1).

¹Justin Evans, Statistique Canada, 170, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6 (justin.evans@statcan.gc.ca); Laura Wile, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6 (laura.wile@statcan.gc.ca)

Tableau 1-1**Exemple de fractionnement entre la version 1.3 de la CNP 2016 et la version 1.0 de la CNP 2021**

CNP ₂₀₁₆	CNP ₂₀₂₁
6541 – Agents/agentes de sécurité et personnel assimilé des services de sécurité	45100 – Surveillants/surveillantes d’élèves, brigadiers/brigadières et autres professions connexes
	64410 – Agents/agentes de sécurité et personnel assimilé des services de sécurité
	65329 – Autre personnel de soutien en service

Il a fallu utiliser un vaste processus pour rajuster les données historiques de l’EPA de la classification de la CNP₂₀₁₆ en fonction de la classification de la CNP₂₀₂₁. En raison du volume de données de l’EPA appartenant aux catégories de la CNP₂₀₁₆ qui ont été fractionnées selon la classification de la CNP₂₀₂₁, des enregistrements d’un échantillon de la période de référence allant de 1987 à 2022 ont été sélectionnés et classés à nouveau de manière manuelle. En fonction des proportions de fractionnement de la CNP₂₀₂₁ observées, toutes les autres périodes de référence ont été estimées au moyen d’une méthode d’interpolation linéaire avant d’être attribuées à l’aide d’une méthode d’allocation aléatoire (AR). Le tout pouvait être examiné par des spécialistes du domaine. Ainsi, les répondants auxquels une catégorie fractionnée de la CNP₂₀₁₆ avait été attribuée ont obtenu, de manière aléatoire, une catégorie de la CNP₂₀₂₁ qui était représentative au fil du temps.

Au milieu de 2023, on s’est aperçu que certains problèmes touchaient la méthodologie utilisée pour rajuster les données historiques de l’EPA en fonction de la classification de la CNP₂₀₂₁. Les utilisateurs des données de l’EPA ont observé des bris de série entre 2023 et les occasions historiques. En outre, en raison des relations incohérentes entre les réponses écrites par les répondants et les catégories attribuées de la CNP₂₀₂₁ aux données historiques, certaines catégories inattendues de la CNP₂₀₂₁ ont été attribuées par le modèle de production de fastText. Depuis 2021, l’EPA utilise des modèles de fastText lors de ses activités de codage manuelles (Evans et Oyarzun, 2021), fournissant des prédictions exactes et uniformes au fil du temps. Cependant, en raison du processus d’allocation aléatoire, le modèle d’apprentissage automatique ne pouvait plus faire de prédictions fiables pour certaines catégories fractionnées de la CNP₂₀₂₁. Ce problème a été exacerbé dans les catégories de la CNP₂₀₁₆ qui affichaient d’importants déséquilibres au chapitre du fractionnement des catégories.

Le présent document résume une nouvelle méthode pour le rajustement de la classification de texte de l’EPA à l’aide de l’apprentissage automatique et étudie ce qui suit :

- L’apprentissage automatique jumelé à une restriction par programmation linéaire (PL), une amélioration méthodologique des pratiques actuelles de rajustement.
- Les critères de sélection de la méthode de rajustement : une évaluation de la méthode proposée pour chaque catégorie fractionnée de la CNP₂₀₁₆.
- De nouveaux modèles : une évaluation des répercussions en aval de la méthode de rajustement proposée sur les modèles de production.

2. Apprentissage automatique

2.1 Algorithme et programmation linéaire

FastText, un algorithme d’apprentissage automatique créé par le laboratoire de recherche en intelligence artificielle de Facebook, a été retenu dans le cadre de la méthode proposée de classification de texte. Lors de la classification de texte à multiples étiquettes, fastText utilise des plongements de mots, qui comprennent des mots mal orthographiés ou inventés, et des mots concaténés, afin de calculer la distribution des probabilités qu’une réponse écrite appartienne à chaque catégorie. FastText applique ensuite une opération d’argument du maximum pour déterminer la catégorie qui correspond à la valeur de probabilité la plus élevée (Joulin et coll., 2016). De façon pratique, fastText appliquera toujours les réponses écrites avec le même texte à la même catégorie, et est plus susceptible d’attribuer des réponses

écrites semblables à une catégorie semblable sur le plan sémantique. C'est exactement le contraire d'une méthode d'allocation aléatoire, dans le cadre de laquelle on ne tient pas compte des réponses écrites.

Pour respecter les proportions observées lors du rajustement, on a utilisé une méthode de programmation linéaire pour restreindre la matrice de probabilité de fastText. Il est possible d'exprimer le problème d'optimisation selon la notation mathématique suivante :

Maximiser :

$$\sum_{i=0}^n \sum_{j=1}^m x_{ij} \cdot p_{ij}$$

Soumis à des contraintes :

$$C1: \sum_{i=1}^n x_{ij} = LC_j \text{ for } j = 1, 2, \dots, m$$

$$C2: \sum_{j=1}^m x_{ij} = 1 \text{ for } i = 1, 2, \dots, n$$

Où :

- n est le nombre d'enregistrements
- m est le nombre d'étiquettes
- x_{ij} est la variable binaire indiquant si l'enregistrement i est attribué à l'étiquette j , où $i = 1, 2, \dots, n$ et $j = 1, 2, \dots, m$
- p_{ij} est la cote de probabilité (entre 0 et 1) associée à l'enregistrement i pour l'étiquette j
- LC_j est le chiffre nécessaire de l'étiquette j .

Tout simplement dit, la méthode cherche à maximiser les cotes de probabilité de fastText, attribuant la catégorie la plus vraisemblable à des réponses écrites données, tout en étant restreinte aux proportions attribuées précédemment. Cela permet de minimiser l'incidence sur la comparabilité des estimations.

2.2 Modélisation

Des modèles ont été créés au moyen des enregistrements fractionnés de la CNP₂₀₁₆ qui avaient été manuellement reclassifiés par le centre d'expertise en codage de Statistique Canada. Selon les auteurs, même si un seul modèle de fastText est utilisé pour la classification de la CNP₂₀₂₁ en production, des modèles distincts pour chacune des 60 catégories fractionnées de la CNP₂₀₁₆ ont enregistré une exactitude globale supérieure (86,8 %) à un modèle unique pour tous les fractionnements (71,7 %). Les deux méthodes d'apprentissage automatique ont dépassé la méthode d'allocation aléatoire de référence (exactitude de 60 %).

3. Critères de sélection de la méthode de rajustement

Pour chacune des 60 catégories fractionnées de la CNP₂₀₁₆, on a comparé trois méthodes différentes (tableau 3-1) :

Tableau 3-1
Méthodes de rajustement envisagées

Méthode	Description
AR	Maintenir les catégories historiques actuelles, attribuées aléatoirement à l'aide de la méthodologie des proportions.
AA	Utiliser un modèle de fastText pour attribuer des prédictions aux enregistrements fractionnés, sans contrainte (attribution de la première prédiction).
AA-PL	Utiliser le modèle de l'apprentissage automatique, mais attribuer des prédictions à l'aide d'une contrainte par programmation linéaire sur une matrice de probabilité de fastText pour respecter les proportions observées au cours du recodage.

3.1 Mesures

Des recommandations ont été faites dans le but d'améliorer le codage au niveau de l'enregistrement tout en tenant compte des bris de série de la CNP₂₀₂₁ apparemment introduits par la méthode d'AR. De nombreuses mesures ont été envisagées (tableau 3.1-1) pour chaque série de la CNP₂₀₁₆. Pour évaluer si l'adoption d'une méthode de rajustement au moyen de l'AA peut avoir des répercussions sur les estimations publiées de l'EPA, on a produit des estimations révisées pour les comparer aux estimations publiées (AR). Des bris de série ont été détectés par des intervalles de confiance sans chevauchement pour deux différentes estimations de l'emploi pour la même CNP₂₀₂₁.

Tableau 3.1-1
Mesures prises formuler une recommandation concernant le rajustement

Méthode	Description
Entre bris de série	Privilégier la méthode (AA ou AA-PL) avec le moins grand nombre de bris (AA/AR ou AA-PL/AR), par rapport à la série d'AR actuellement à la disposition des utilisateurs de l'EPA.
Au sein de bris de série : occasions historiques	Privilégier la méthode (AA, AA-PL ou AR) avec le moins grand nombre de bris d'une année à l'autre et d'un mois à l'autre, sauf si les bris sont prévus.
Modification de la proportion d'enregistrements	Privilégier l'AA si une proportion similaire d'enregistrements a été modifiée par l'AA ou l'AA-PL par rapport à l'AR.
Taux de correspondance avec le Système de codage selon les classifications (SCC) de Statistique Canada	Le SCC comprend des milliers de titres de poste et la catégorie de la CNP ₂₀₂₁ associée. Privilégier la méthode (AA, AA-PL ou AR) avec le taux de correspondance le plus élevé lors d'une comparaison avec le SCC.
Consultation avec des experts en codage	Pour 11 des 60 catégories de la CNP ₂₀₁₆ , des codeurs experts de l'EPA ont passé en revue les catégories de la CNP ₂₀₂₁ attribuées et ont fourni une recommandation.

Tableau 3.1-2
Estimations de la CNP₂₀₂₁ à l'échelle nationale pour les 10 dernières années (de 201301 à 202306)

Niveau de la CNP	Nombre de catégories	Au moins 1 bris de série (/126 occasions, IC de 95 %)	Bris en pourcentage
1	9	0	0,00 %
2	33	1	3,03 %
3	55	1	1,82 %
4	71	3	4,23 %
5	116	11	9,48 %

Les bris de série ont été classés en fonction de l'incidence sur la volatilité à court terme (nombre de bris entre 2022 et 2023) et la volatilité à long terme (nombre de bris d'une année sur l'autre entre 2013 et 2022). À l'échelle nationale,

il y avait 11 catégories de la CNP-5 sur 116 qui comptaient au moins un bris de série (tableau 3.1-2). On passait à 4 catégories de la CNP-5 sur 116 avec au moins 5 mois et affichant une différence significative entre la série proposée et la série d'AR. Cependant, cette situation peut être attribuable à un faible nombre d'enregistrements ou être considérée comme étant justifiable si l'on fait une comparaison avec les tendances observées en 2023. Au niveau provincial, les bris de séries de la CNP₂₀₂₁ ont été principalement attribués à un CNP-5 ou à un faible nombre d'enregistrements. Dans les séries, seulement 1 CNP-5 affichait des bris remarquables d'un mois à l'autre. Cependant, cette situation s'expliquait par la saisonnalité attendue qui était absente pour certaines occasions de la série d'AR. Dans le cas des séries pour lesquelles un nombre supérieur de bris d'une année sur l'autre ont été observés, les auteurs ont consulté les codeurs spécialisés de l'EPA pour formuler une recommandation. Ensemble, ces résultats suggèrent que les méthodes proposées auront des répercussions minimales sur les estimations publiées.

3.2 Comparaison des réponses écrites

Le Système de codage selon les classifications (SCC) est un outil interne de Statistique Canada utilisé pour aider les codeurs manuels du centre d'expertise à appairer les réponses écrites des répondants avec une catégorie de classification. Dans l'exemple hypothétique ci-dessus (tableau 3.2-1), la réponse écrite « brigadier scolaire », attribuée historiquement à la CNP₂₀₁₆6541, pourrait être rajustée pour être attribuée à la CNP₂₀₂₁64410 à l'aide de la méthode d'AR, car cette méthode ne tient pas compte des réponses écrites. De manière générale, une prédiction de l'AA correspondait plus souvent au SCC qu'à une catégorie attribuée par la méthode d'AR.

Tableau 3.2-1

Comparaison hypothétique des réponses écrites pour la CNP₂₀₁₆6541

Texte du Système de codage selon les classifications	Catégorie de la CNP ₂₀₂₁ du Système de codage selon les classifications	Catégorie de la CNP ₂₀₂₁ de l'allocation aléatoire (AR)	Catégorie de la CNP ₂₀₂₁ de l'apprentissage automatique (AA)
Brigadier scolaire	45100 – Surveillants/surveillantes d'élèves, brigadiers/brigadières et autres professions connexes	64410 – Agents/agentes de sécurité et personnel assimilé des services de sécurité	45100 – Surveillants/surveillantes d'élèves, brigadiers/brigadières et autres professions connexes
Agent de sécurité	64410 – Agents/agentes de sécurité et personnel assimilé des services de sécurité	45100 – Surveillants/surveillantes d'élèves, brigadiers/brigadières et autres professions connexes	64410 – Agents/agentes de sécurité et personnel assimilé des services de sécurité

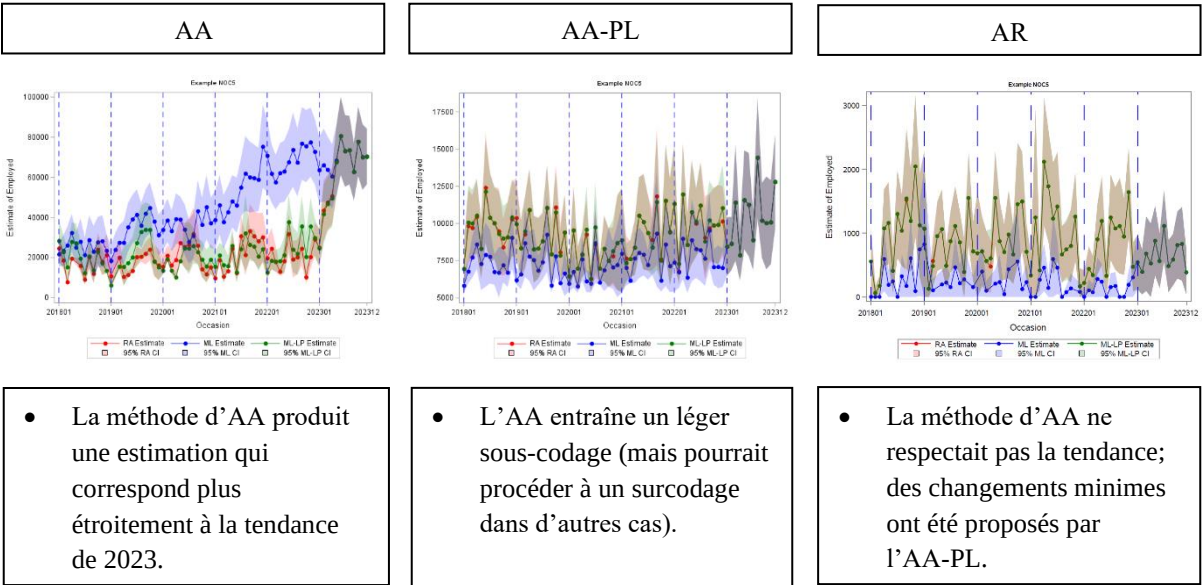
3.3 Recommandations

Les méthodes proposées ont toutes été utiles pour assurer un rajustement en fonction de conditions particulières (figure 3.3-1). Les auteurs ont recommandé, pour la majorité des catégories fractionnées de la CNP₂₀₁₆, d'utiliser une méthode de rajustement au moyen de l'AA avec ou sans contrainte (tableau 3.3-1). Malgré sa capacité à harmoniser des réponses écrites semblables au niveau des enregistrements, l'apprentissage automatique (méthode d'AA) avait la propension de procéder à un sous-codage ou à un surcodage d'une série à fort déséquilibre de catégories. Dans ce cas-ci, les restrictions de proportion (méthode d'AA-PL) ont permis de procéder à un rajustement correctif tout en respectant les estimations publiées. Enfin, dans les cas où la méthode AA ou AA-PL ne fournissait pas d'amélioration suffisante, la série a été laissée inchangée (méthode d'AR).

Tableau 3.3-1
Résumé de la décision concernant le fractionnement de la CNP₂₀₁₆
Méthode recommandée Chiffres de la CNP₂₀₁₆

AA	23
AA-PL	31
AR	6
Total	60

Figure 3.3-1
Exemples représentatifs de séries chronologiques



4. Nouvelle modélisation

En fonction des décisions relatives au fractionnement de la CNP₂₀₁₆, on a établi un nouveau modèle à l'aide de données historiques recodées selon les catégories recommandées de la CNP₂₀₂₁, afin de faire une simulation du modèle de production qui serait créé à l'aide de la méthodologie proposée. Le modèle a été établi au moyen des mêmes ensembles de données d'entraînement et d'essai utilisés dans le modèle de production actuel (avec les catégories mises à jour de la CNP₂₀₂₁) et validés pour les données de l'EPA des 6 mois les plus récents. Dans ce cas-ci, on peut voir que, même si les deux modèles affichent un degré d'exactitude globale semblable (73,8 % contre 74,9 %, sans utiliser de seuil), dans le cas des enregistrements fractionnés, le modèle rajusté a affiché un meilleur rendement (+5,3 points de pourcentage) que le modèle de production (tableau 4-1). Ensuite, on a imposé une exigence relative à la précision par catégorie à multiples seuils pour déterminer le nombre de catégories à utiliser dans un scénario de production. Comme prévu, le modèle rajusté comportait un nombre supérieur de catégories fractionnées (+24) avec un taux de précision d'au moins 85 %.

Tableau 4-1
Comparaison des modèles pour un ensemble de données d’essai d’un an

	Mesures	Modèle de production	Modèle mis à jour
Ensemble de données d’essai Seuils multiples	Exactitude globale	73,8 %	74,9 %
	Exactitude de fractionnement	66,2 %	71,5 %
	Catégories avec exigence en matière de précision de 85 %	401	439
	Catégories fractionnées avec exigence en matière de précision de 85 %	67	101

Selon la méthode d’allocation aléatoire, les catégories de la CNP₂₀₂₁ sont moins susceptibles de correspondre aux réponses écrites lors de fractionnements de faible proportion. Lors de l’examen de l’exactitude des catégories, le modèle de production actuel, établi à l’aide de données de l’allocation aléatoire, affiche le pire rendement dans le cas des catégories associées à des proportions de fractionnement faibles (tableau 4-2). En revanche, les catégories comportant une proportion de fractionnement élevée affichent une exactitude de modèle relativement semblable. Cela souligne l’importance des activités de rajustement pour assurer la qualité des données, car cela peut avoir une incidence sur les décisions relatives au modèle en aval.

Tableau 4-2
Exactitude des catégories fractionnées dans les modèles pour un ensemble de données de validation de 6 mois

CNP ₂₀₂₁	Proportion de fractionnement	Enregistrements	Exactitude du modèle de production (%)	Exactitude du modèle mis à jour (%)	Différence
Catégorie 1	0,02	38	0 %	84 %	+84 %
Catégorie 2	0,08	28	0 %	75 %	+75 %
...	...	...	...	...	...
Catégorie 115	0,98	13	76 %	69 %	-8 %
Catégorie 116	0,56	24	25 %	17 %	-8 %

5. Résumé

On a pu constater que la méthodologie de rajustement au moyen de l’AA représente de nouvelles tendances, minimise les bris de séries chronologiques et harmonise les réponses écrites historiques. Selon les auteurs, les méthodes proposées améliorent les données historiques de l’EPA, alimentant les prédictions de l’AA en aval qui doivent être de la qualité nécessaire afin de produire des statistiques officielles.

Remerciements

Les auteurs souhaitent remercier les personnes suivantes pour leurs contributions : Nadya Aliksandrovich, Erin Howlett, David Menyah, Brittany Milton, Eric Olson, Sylvia White, Cynthia Breault, Alexander Calvano, Benoit Durand, Brandon Gould, Johanne Madore, Derek Strelloff, Eric Laflamme, Julie Portelance, Kerry Schneider, Jessica Mulligan, Javier Oyarzun, Michelle Simard, et Yi Li.

Bibliographie

- Evans, J.J., et Oyarzun, J. (2021), “Need for Speed: Using fastText (Machine Learning) to Code the Labour Force Survey”, *Proceedings of Statistics Canada Symposium 2021*. Accessible à l’adresse <https://www150.statcan.gc.ca/n1/en/catalogue/11-522-X202100100013>
- Statistics Canada (2023), “Introduction to the National Occupational Classification (NOC) 2021 Version 1.0”, Extrait de <https://www.statcan.gc.ca/en/subjects/standard/noc/2021/introductionV1#a8>.
- Joulin, A., Grave, E., Bojanowski, P., et Mikolov, T. (2016), “Bag of Tricks for Efficient Text Classification”, <https://arxiv.org/pdf/1607.01759>.