Recueil du Symposium 2018 de Statistique Canada
Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données

Le jeu de l'imitation : Un aperçu d'une approche de l'apprentissage automatique pour une codification de la classification des industries

Javier Oyarzun[1]

Résumé

Le Registre des entreprises (RE) de Statistique Canada joue un rôle fondamental dans le mandat de Statistique Canada. Le RE est un répertoire qui indiquent toutes les entreprises exploitées au Canada. Près de 200 enquêtes-entreprises recourent au RE de diverses façons. Le RE a une incidence directe sur l'efficience du processus d'enquêtes-entreprises, sur la fiabilité des données produites par les programmes de statistiques des entreprises et sur la cohérence du Système de comptabilité nationale. L'un de ses attributs clés est le code industriel. Au début de 2018, Statistique Canada a commencé à élaborer une nouvelle méthodologie pour coder de façon probabiliste la classification industrielle des entreprises. Cette méthodologie, qui utilise l'extraction à partir de textes et l'apprentissage automatique, fournira à Statistique Canada un outil pour coder les classifications industrielles manquantes et améliorer la qualité globale des classifications industrielles dans le RE.

Le présent article fournit une explication sur le Système de classification des industries de l'Amérique du Nord (SCIAN), son utilisation dans les programmes statistiques de Statistique Canada, les méthodes actuelles et nouvelles de codification du SCIAN, ainsi qu'une discussion sur les défis liés aux entreprises non codifiées et des exemples de cas complexes de codification du SCIAN.

Mots-clés : Registre des entreprises; apprentissage automatique; extraction à partir de textes; SCIAN; codification; industriel; classification.

1. Introduction

Le Registre des entreprises (RE) de Statistique Canada joue un rôle fondamental dans le mandat de Statistique Canada. Le RE est un répertoire qui indiquent toutes les entreprises exploitées au Canada. Près de 200 enquêtes-entreprises utilisent le RE de différentes façons, principalement pour établir une base de sondage, pour prélever un échantillon, pour recueillir et traiter des données et pour établir des estimations. Le RE a une incidence directe sur l'efficience du processus d'enquêtes-entreprises, sur la fiabilité des données produites par les programmes de statistiques des entreprises et sur la cohérence du Système de comptabilité nationale. Une telle entreprise exige un processus fiable de contrôle de la qualité. L'un de ses attributs clés est le code industriel du Système de classification des industries de l'Amérique du Nord (SCIAN) attribué à chaque entreprise. Le SCIAN est un code à six chiffres attribué en fonction de la principale activité commerciale d'une entreprise donnée. À l'heure actuelle, environ 500 000 entreprises, sur le total des entreprises actives, lesquelles sont au nombre d'environ 7 000 000, demeurent non codifiées, la vaste majorité d'entre elles étant des entreprises relativement petites. Au début de l'année 2018, Statistique Canada a commencé à développer des algorithmes d'apprentissage automatique pour codifier ces unités afin d'accroître la couverture de la classification des entreprises dans le but d'améliorer les estimations économiques de l'ensemble des programmes d'enquête économique.

Le présent article est divisé en cinq sections et examine les multiples facettes de cette nouvelle méthodologie de codification. La première section offre une brève introduction au RE. La deuxième est une discussion sur le SCIAN et son utilisation à Statistique Canada. La troisième présente les méthodes de codification du SCIAN adoptées par le RE, et la quatrième présente les méthodes de codification du SCIAN que sont l'extraction à partir de textes et l'apprentissage automatique. Enfin, dans la dernière section présente une discussion sur la population des entités non codifiées du RE (« backlog ») et l'application des nouvelles méthodes de codification du SCIAN.

[1] Javier Oyarzun, Statistique Canada, 100 promenade du pré Tunney, Ottawa (Ontario), Canada, K1A 0T6 (javier.oyarzun@canada.ca)

2. SCIAN

Le Système de classification des industries de l'Amérique du Nord (SCIAN) est un système de classification des industries qui a été conçu par les organismes statistiques du Canada, du Mexique et des États-Unis. Créé avec comme toile de fond l'Accord de libre-échange nord-américain, le SCIAN vise à fournir des définitions communes de la structure industrielle des trois pays, ainsi qu'un cadre statistique commun pour faciliter l'analyse des trois économies. Le SCIAN est articulé autour des principes de l'offre ou de la production, afin de s'assurer que les données sur les industries qui sont classées en fonction du SCIAN se prêtent à l'analyse de questions liées à la production, comme le rendement industriel[2].

Le SCIAN a été conçu pour classifier les entreprises et les autres organismes qui sont engagés dans la production de biens et de services. Il s'agit des entreprises constituées en sociétés (sociétés, T2), et des entreprises non constituées en sociétés (T1). Elles comprennent également les institutions et les organismes publics qui fournissent des services marchands et non marchands ainsi que des organisations telles que les associations professionnelles, les syndicats, les organismes de bienfaisance ou sans but lucratif et les employés des ménages. Le SCIAN est un code complet à six chiffres parmi 928 options attribué à chaque entreprise en fonction de sa principale activité. La structure du SCIAN est hiérarchique. Elle comprend des secteurs (codes à deux chiffres), des sous-secteurs (codes à trois chiffres), des groupes d'industries (codes à quatre chiffres) et des industries (codes à cinq chiffres). Le sixième chiffre sert à désigner les industries nationales. Le tableau 2-11 présente la classification du SCIAN au niveau des deux chiffres.

Tableau 2-1
Niveaux SCIAN à deux chiffres

SCIAN	Secteur
11	Agriculture, foresterie, pêche et chasse
21	Extraction minière, exploitation en carrière, et extraction de pétrole et de gaz
22	Services publics
23	Construction
31-33	Fabrication
41	Commerce de gros
44-45	Commerce de détail
48-49	Transport et entreposage
51	Industrie de l'information et industrie culturelle
52	Finance et assurances
53	Services immobiliers et services de location à bail
54	Services professionnels, scientifiques et techniques
55	Gestion de sociétés et d'entreprises
56	Services administratifs, services de soutien, services de gestion des déchets et services d'assainissement
61	Services d'enseignement
62	Soins de santé et assistance sociale
71	Arts, spectacles et loisirs
72	Services d'hébergement et de restauration
81	Autres services (sauf les administrations publiques)
91	Administrations publiques

3. Méthodes actuelles de codification du SCIAN

Le Registre des entreprises est chargé de codifier et de tenir le SCIAN pour l'ensemble des entreprises canadiennes. L'attribution d'un SCIAN aux entreprises canadiennes représente un défi depuis sa création. Au fil des ans, le RE a examiné différentes méthodes pour la codification du SCIAN. Les principales méthodes utilisées en ce moment par le RE sont les suivantes :

[2] Statistique Canada (2018), *La version 3.0 de 2017 du SCIAN.*.

- **Sources administratives – provenant principalement de l'Agence du revenu du Canada (ARC)**
  - Classification des industries autocodée des entreprises (SCIAN)
  - Descriptions de l'activité commerciale (telle que fournie dans les déclarations de revenus des entreprises) – celles-ci sont ensuite traitées au moyen du G-Code de Statistique Canada[3] en fonction du fichier de référence des termes industriels communs.
- **Mises à jour consécutives aux enquêtes**
  - Les entreprises sont sondées ou une prise de contact préalable est effectuée.
- **Activités d'établissement de profil**
  - Communication périodique avec les entreprises afin d'en établir le profil pour la tenue du RE.

Ces méthodes comportent des forces et des faiblesses. Par exemple, le SCIAN autocodé a été critiqué pour sa mauvaise qualité dans certains secteurs industriels. Or, il s'agit souvent de la seule source d'information accessible (surtout pour les T1). La codification fondée sur le champ de description des activités de l'ARC au moyen de G-Code nécessite que la description de l'activité commerciale soit indiquée, ce qui n'est pas toujours le cas. La codification au moyen d'enquêtes est coûteuse et accroît le fardeau de réponse. La codification manuelle, au moyen d'enquêtes ou de l'établissement de profils, peut améliorer la qualité, mais nécessite beaucoup de temps et de formation.

4. Nouvelles méthodes de codification du SCIAN

Au début de l'année 2018, Statistique Canada a décidé d'examiner de nouvelles méthodes de codification du SCIAN non seulement pour améliorer la qualité de la codification du SCIAN dans le RE, mais aussi pour attribuer un SCIAN aux entités non codées du RE (consulter la section 5 pour obtenir de plus amples renseignements). Les méthodes suivantes feront l'objet d'une discussion plus détaillée dans les sous-sections ci-après : (4.1) Extraction à partir de textes et (4.2) Apprentissage automatique.

4.1 Extraction à partir de textes

L'extraction à partir de textes est le processus qui consiste à dériver des renseignements de grande qualité à partir de textes. L'extraction à partir de textes fait habituellement intervenir l'extraction d'information, le traitement préalable pour convertir les données d'entrée textuelles brutes dans un format structuré, le repérage de tendances dans les données structurées dérivées qui en résultent, l'analyse lexicale pour étudier les distributions de fréquence des mots, la reconnaissance de tendances et l'analytique prédictive. Une application typique consiste à numériser un ensemble de documents (communément appelé « corpus » selon la terminologie de l'extraction à partir de textes) rédigés en langage naturel et soit à modéliser le corpus aux fins de classification prédictive ou à alimenter un certain index de recherche au moyen de l'information extraite.

4.1.1 Codification de texte

Pour chaque mot propre au SCIAN dans le nom d'une entreprise, le champ de description d'industrie ou le champ de la principale activité peut être prédominant (ou même unique). Par exemple, le mot « BRDCST » apparaît dans les descriptions d'industrie de plus de 6 000 entreprises, lesquelles appartiennent toutes à la classification du SCIAN 519130 (Édition, radiodiffusion et télédiffusion par Internet et sites portails de recherche). Par contraste, le mot plus commun « ŒUFS » apparaît dans les descriptions d'industrie de plus de 20 000 entreprises réparties sur plusieurs codes du SCIAN. Parmi les entreprises dont les descriptions d'industrie contiennent le mot « ŒUFS », le SCIAN le plus commun est 112310 (Production d'œufs de poules), qui survient 30,4 % du temps, alors que le SCIAN 413130 (Grossistes-marchands de volailles et d'œufs) survient environ 20 % du temps parmi ces entreprises. Le tableau 4.1.1-1 indique un certain nombre d'exemples de mots dans les descriptions d'industrie, le SCIAN le plus commun parmi les entreprises dont les descriptions d'industrie contiennent ces mots :

[3] Statistics Canada, (2018), « *G-Code version 3.0* ».

$$P(SCIAN|W_{SCIAN}) = \frac{Freq(W_{SCIAN})}{\sum_{k \in tout\ SCIAN} Freq(W_k)}$$

où $Freq(W_{SCIAN})$ est le nombre d'entreprises dont les descriptions d'industrie contiennent le mot *W* et dont la classification SCIAN est SCIAN.

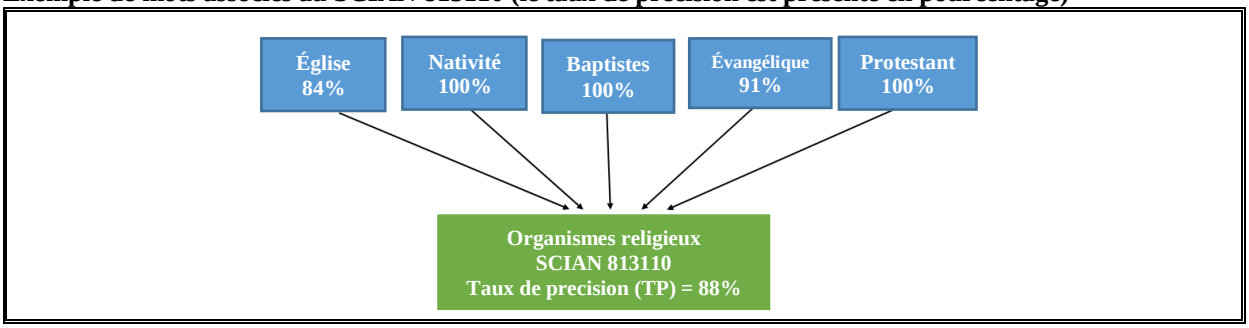
Tableau 4.1.1-1
Mot dans le champ de description et le SCIAN qui y est associé

Mot (W)	Freq(W))	P(SCIAN|W_SCIAN)	SCIAN	Titre dans le SCIAN
BRDCST	6 000	100,0 %	519130	Édition, radiodiffusion et télédiffusion par Internet et sites portails de recherche
SONDAGE	4 000	100,0 %	541910	Études de marché et sondages d'opinion
APICULTURE	2 000	100,0 %	112910	Apiculture
UNIFAMILIALES	1 000	100,0 %	236110	Construction résidentielle
BEIGNE	400	100,0 %	722512	Établissements de restauration à service restreint
BAPTISTE	200	100,0 %	813110	Organismes religieux
CAMPINGS	150	100,0 %	721211	Parcs pour véhicules récréatifs (VR) et camps de loisirs
SERVICES DE CONCIERGE	9 000	88,9 %	561722	Services de conciergerie, sauf le nettoyage de vitres
ÉGLISE	15 000	84,1 %	813110	Organismes religieux
RÉCLAMATION ASSURANCE	200	75,6 %	524291	Experts en sinistres
ŒUFS	20 000+	30,4 %	112310	Production d'œufs de poules

4.1.2 Prédire le SCIAN à l'aide de l'extraction à partir de textes

Comme il est indiqué dans la sous-section 4.1.1., il est possible de prédire le SCIAN d'une entreprise en fonction des mots et des noms au moyen d'algorithmes d'extraction à partir de textes. Donc, différents mots ou noms peuvent être utilisés pour prédire un SCIAN précis. Par exemple, des mots comme « église », « nativité », « baptistes », et de nombreux autres peuvent aider à coder des entreprises dans le SCIAN 813110 (Organismes religieux) : voir la figure 4.1.2-1.

Figure 4.1.2-1
Exemple de mots associés au SCIAN 813110 (le taux de précision est présenté en pourcentage)



Le tableau 4.1.2-1 présente le taux de précision atteint lorsqu'on utilise l'extraction à partir de textes au moyen des noms d'entreprises (NM), du champ de description (DM) et du champ de la principale activité (AM). On a constaté que la codification du SCIAN était de meilleure qualité lorsque le champ de description était utilisé. Cela signifie que les mots dans le DM présentent une association plus forte au SCIAN que les mots présents dans les NM ou l'AM. Le tableau indique également que le taux de précision est beaucoup plus faible pour les secteurs manufacturiers (SCIAN-

2 : 31, 32 et 33), le commerce de gros (SCIAN-2 : 41) et la gestion (SCIAN-2 : 55), indépendamment de l'utilisation des NM, du DM ou de l'AM. Par conséquent, les mots que l'on trouve dans les entreprises de ces secteurs ne sont pas de bons prédicteurs du SCIAN.

Tableau 4.1.2-1
Taux de précision (%) atteint lorsqu'on utilise les noms d'entreprises (NM), le champ de description (DM) et le champ de la principale activité (AM)

SCIAN	11	21	22	23	31-33	41	44-45	48-49	51	52	53	54	55	56	61	62	71	72	81	91	Total
NM	63	49	54	90	36	44	85	83	65	80	91	80	2	76	29	91	58	95	78	82	83
DM	98	80	86	89	66	24	90	97	91	97	99	93	75	93	94	98	97	96	90	88	94
AM	95	39	24	94	31	13	79	78	71	90	39	92	2	82	60	98	77	97	85	93	87

4.2 Apprentissage automatique

Après la codification du SCIAN et l'extraction à partir de textes, le RE a décidé de se pencher sur l'utilisation d'un classifieur bayésien naïf de Bernoulli multivarié pour prédire le SCIAN, en fonction des caractéristiques dérivées des variables textuelles d'entrées ci-après :

- Nom de l'entreprise
- Champ de description de l'entreprise
- Champ de la principale activité

Le nom, la description et la principale activité de chaque entreprise ont été concaténés et traités comme un seul document. Un vocabulaire a été construit en fonction des mots qui surviennent dans le corpus de documents qui en résultent. Un vecteur multivarié à une seule caractéristique a ensuite été généré pour chaque document en fonction de la fréquence des mots dans le document donné. Les résultats étaient de meilleure qualité que ceux de la méthode de l'extraction à partir de textes présentés à la section 4.1; voir le tableau 4.2.1-2. Voici les étapes utilisées dans le déroulement des opérations du traitement préalable, de l'élaboration de caractéristiques et de l'apprentissage automatique.

- Normalisation des mots : majuscules, supprimer les accents, indexation sur le début du mot, etc.
- Établissement d'une matrice sur la fréquence dans les documents (MFD) : Utilisation possible des n-grammes
- Choix des paramètres idéaux dans la MFD : Exactitude la plus élevée avec les données d'essai et de formation
- Établissement du modèle
- Prédiction
- Évaluation et prise de décision

4.2.1 Classifieurs bayésiens naïfs

Les classifieurs bayésiens naïfs (NB) sont une collection de techniques de classification dont les modèles sous-jacents sont probabilistes et adoptent l'hypothèse bayésienne naïve. Selon l'hypothèse bayésienne naïve, les caractéristiques sont conditionnellement indépendantes compte tenu de l'étiquette de classe. Pour certains problèmes de classification, les classifieurs bayésiens naïfs peuvent être formés très efficacement. Pour cette raison, les classifieurs bayésiens naïfs demeurent des méthodes prisées pour la catégorisation textuelle (voir les exemples du tableau 4.2.1-1). Aux fins du codification de l'apprentissage automatique du SCIAN, le paquet R *Quanteda* a été utilisé, lequel fournit la mise en œuvre du classifieur bayésien naïf de Bernoulli multivarié. Le tableau 4.2.1.2 présente le taux de précision de cette méthode en fonction des secteurs industriels. La méthode du classifieur bayésien naïf utilise la formule ci-après :

$$\boldsymbol{P(C|X_i)} = \frac{\prod_i \big(\boldsymbol{P(X_i|C)}\big)\, \boldsymbol{P(C)}}{\boldsymbol{P(X_i)}}$$

Où C est l'étiquette de classe (SCIAN) et X_i sont les caractéristiques (mots).

$P(C)$ est la probabilité que l'entreprise appartienne à un code SCIAN précis.
$P(X_i|C)$ est la probabilité des entreprises comportant une combinaison de caractéristiques précises, sachant qu'elle appartient à un SCIAN précis.

Tableau 4.2.1-1
Codification du SCIAN par l'apprentissage automatique bayésien naïf, en fonction des mots

Word	NAICS 1	Title 1	Prob 1	NAICS 2	Title 2	Prob 2	NAICS 3	Title 3	Prob 3
COURTIERS	524291	Experts en sinistres	75,6 %	524210	Agences et courtiers d'assurance	12,3 %	524299	Toutes les autres activités liées à l'assurance	3,4 %
COURTIERS D'ASSURANCE	524291	Experts en sinistres	88,0 %	524210	Agences et courtiers d'assurance	6,8 %	524299	Toutes les autres activités liées à l'assurance	3,0 %
COMPAGNIE « H & S »	524291	Experts en sinistres	98,2 %	524210	Agences et courtiers d'assurance	1,4 %	524299	Toutes les autres activités liées à l'assurance	0,2 %

Tableau 4.2.1-2
Taux de précision du SCIAN codé au moyen de l'apprentissage automatique (%)

SCIAN	11	21	22	23	31-33	41	44-45	48-49	51	52	53	54	55	56	61	62	71	72	81	91	Total
NB (1)	99	94	96	99	88	96	95	98	94	98	100	98	80	97	98	98	99	95	96	99	97
NB (2)	93	83	78	85	72	75	91	88	89	88	96	85	77	78	88	92	93	90	84	86	89

Notes : (1) Classifieur bayésien naïf avec un seul de probabilité a posteriori supérieur à 80 %. (2) Classifieur bayésien naïf sans seuil de probabilité.

5. Entreprises non codées du Registre des entreprises

Le terme « backlog » renvoie à la population « active » actuellement inscrite au RE que les méthodes actuelles de codification du SCIAN ne parviennent pas à coder. Ces unités ne peuvent donc pas faire partie du champ de portée des enquêtes économiques axées sur le SCIAN (soit la majorité des enquêtes-entreprises de Statistique Canada), ce qui peut donner lieu à la sous-estimation des paramètres économiques. Selon les principales mesures de la taille, le « backlog » représente :

- 500 000 établissements actifs (de 5 % à 8 % du RE);
- 200 milliards de dollars en revenus (de 3 % à 5 % du RE);
- 400 000 dans le secteur de l'emploi (de 2 % à 3 % du RE).

Comme il est indiqué dans la section 4, les techniques d'extraction à partir de textes et l'apprentissage automatique pourraient être utilisés pour coder un nombre considérable d'unités du « backlog ». Les tableaux 4.1.2-1 et 4.2.1-2 indiquent le taux de précision des différentes techniques d'extraction à partir de textes et de l'apprentissage automatique (extraction des noms d'entreprises à partir des textes [NM], extraction des descriptions à partir des textes [DM] et extraction des activités fiscales à partir des textes [AM] et classifieur bayésien naïf [NB]). Le tableau 5-1 présente le nombre d'unités pouvant être codées au moyen des méthodes mentionnées plus tôt.

Tableau 5-1

« Backlog » codé au moyen de l'extraction à partir de textes et de l'apprentissage automatique

SCIAN	NM	DM	AM	NB
11	441	3 790	312	4 997
21	100	141	27	2 757
22	14	498	1	224
23	4 035	10 847	2 145	21 327
31-33	195	2 002	42	1 020
41	335	505	14	7 347
44-45	2 083	4 164	1 079	8 150
48-49	636	5 754	268	6 787
51	533	4 038	778	3 200
52	1 762	11 980	782	32 073
53	5 705	5 892	138	28 650
54	2 959	19 411	5 605	42 877
55	9	362	0	30 096
56	1 827	8 596	603	10 288
61	204	1 504	463	3 321
62	3 742	5 408	2 278	8 095
71	337	4 447	603	4 569
72	2 730	3 539	1 341	6 188
81	11 857	10 439	2 865	33 255
91	51	119	15	78
Total	39 555	103 436	19 359	255 299

Conclusion

La précision du RE a une incidence directe sur l'efficience du processus d'enquêtes-entreprises, sur la fiabilité des données produites par les programmes de statistiques des entreprises et sur la cohérence du Système de comptabilité nationale. Les méthodes de l'extraction à partir de textes et de l'apprentissage automatique présentées dans cet article fournissent à Statistique Canada de nouvelles méthodes de haute qualité pour la codification du SCIAN. Ces nouvelles méthodes peuvent être utilisées pour coder une grande partie des entreprises non codées du RE en matière de SCIAN, et avoir une incidence positive sur les estimations du programme économique de Statistique Canada.

Remerciements

L'auteur souhaite remercier les personnes suivantes, qui ont contribué à rendre ce projet possible : Sonja Simic, Aaron McBride, Shuai Zhang, Yi Li, Laura Wile, Kenneth Chu, Christian Wolfe, Anthony Yeung, Danielle Lebrasseur, Alain Therrien, Jamie Brunet, Jeff Mondoux, Linda Scantland, Amanda Maddicks et Stan Hatko.

Bibliographie

Benoit, K., K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, et A. Matsuo. (2018), « Quanteda: An R package for the Quantitative Analysis of Textual Data », *Journal of Open Source Software*, 3(30), p. 774.

Horwood, J. (2018), « Machine Learning for Scanner Data Classification », rapport non publié, Ottawa, Canada: Statistics Canada.

Oyarzun, J. (2018), « NAICS coding: How can we do better? » présenté au Comité technique de la Division des méthodes d'enquêtes auprès des entreprises de Statistique Canada, 25 mai 2018.

Simic, S., et J. Oyarzun (2018), « OK Computer: Using Machine Learning to Code NAICS on the Business Register » présenté aux séminaires de la Division des méthodes d'enquêtes auprès des entreprises de Statistique Canada, 14 juin 2018.

Statistique Canada (2018), *La version 3.0 de 2017 du SCIAN*.