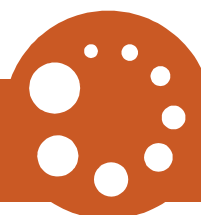




NOGAuto: Automatisation de la codification des unités dans le Registre des entreprises et établissements

Projet pilote de la stratégie d'innovation des données:
Utilisation de technologies d'apprentissage automatique
(Machine Learning)

EXPERIMENTAL STATISTICS



Neuchâtel, 2019

Éditeur: Office fédéral de la statistique (OFS)
Rédaction: Claude Macchi
Domaine: 00 Bases statistiques des généralités
Langue du texte original: Français
Traduction: Services linguistiques de l'OFS

Concept de mise en page: Section DIAM
Téléchargement: www.statistique.ch
Copyright: OFS, Neuchâtel 2019
La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée

Sommaire

1	Objet de ce document	4
2	Résumé	4
3	Situation initiale	4
4	Objectifs	5
5	Résultats	6
5.1	Les éléments de NOGAuto et leur construction	6
5.2	Les premiers résultats	8
6	Ressources	8
7	Planification et organisation	9
8	Conséquences/Lessons learned	9
9	Annexe	11
9.1	Annexe 1 : Résultats des tests effectués selon les méthodes forêt aléatoire et réseau de neurones	11

1 Objet de ce document

Ce document a pour objet de résumer les événements et les résultats du projet « NOGAuto ». Il contient les conclusions qui peuvent être tirées du travail de projet dans le but de faciliter des révisions et des projets futurs.

2 Résumé

Les nomenclatures sont des éléments de base pour la production de statistiques. La qualité du codage des unités observées a un impact direct sur l'ensemble du processus de production des données, sur la crédibilité et sur la qualité du résultat statistique. C'est d'autant plus important dans le contexte des registres et des données administratives, qui sont le point de départ d'innombrables statistiques, la base des cadres de sondage et de l'analyse des données.

Afin d'améliorer en permanence la qualité de la codification des unités dans le SBER (« Swiss Business and Entreprise Register ») et de réduire la charge pesant sur les entreprises dans leurs obligations d'information des offices statistiques, la section URD de l'OFS a lancé, dans le cadre du programme de la stratégie d'innovation sur les données de l'OFS, un projet visant à automatiser l'attribution du code des activités économiques aux entreprises.

Les procédures de codage sont actuellement assez standardisées. Les codeurs analysent et interprètent les informations sur les activités des entreprises, telles que les données fournies par les entreprises elles-mêmes, les résultats d'enquêtes ainsi que les descriptions figurant dans les registres des sociétés et différentes données administratives. Sur cette base ils définissent des mots-clés qui sont comparés à une liste de mots-clés et de concepts liés à chacune des positions de la classification et à leurs notes explicatives et choisissent ainsi le code à attribuer à l'entreprise observée.

Grâce à de nouvelles méthodes d'innovation, la section URD envisage de mettre au point un système d'apprentissage automatique pour automatiser la procédure de codage manuel. Ce système (« NOGAuto ») effectue les étapes de lecture et d'interprétation des descriptions des activités économiques et associe automatiquement l'entreprise à un code de classification. Il s'agit en fait d'apprendre à partir d'un ensemble de données existant qui a déjà été testé manuellement. Par la suite, l'horizon du système pourrait être élargi en cherchant directement sur le Web d'autres sources d'information sur l'entreprise à coder.

Les résultats atteints dans cette première phase du projet sont très prometteurs. Bien qu'avec des ressources très limitées et une infrastructure encore à améliorer, dans le cadre des tests effectués les codes corrects proposés par le système ont atteint 89%.

La continuation du projet et la réalisation d'un système qui puisse assurer une attribution automatique des codes est étroitement relié (1) à l'augmentation des ressources mises à disposition du projet ainsi (2) qu'à l'existence d'une structure qui puisse garantir un suivi technique et une infrastructure à la hauteur des besoins d'un tel projet.

Il est de plus important souligner que les efforts investis pour la réalisation de NOGAuto ne se limiteraient pas seulement à la codification des activités économiques des entreprises, mais qu'avec des petites adaptations le système peut être utilisé aussi pour la codification des professions, des maladies et causes de mort, des produits etc.

3 Situation initiale

Chaque entreprise et établissement stocké dans le SBER dispose d'un code de la nomenclature des activités économiques NOGA. La codification s'effectue en deux étapes principales : la première, avec l'attribution d'un code provisoire lors de l'intégration de l'entreprise dans les registres, et la deuxième,

qui a lieu quelques mois plus tard, avec la validation du premier code attribué par le biais d'une enquête.

La première attribution du code d'activité économique peut être faite soit par l'OFS (principalement dans le cas de données administratives) soit directement par des sociétés tierces (surtout dans le cas d'annonces provenant de registres du commerce).

Le premier code est ensuite validé par l'OFS dans le cadre des enquêtes « ERST » ou « Profiling Light ». Sur la base des descriptions des activités économiques fournies par les entreprises elles-mêmes, les codeurs identifient les termes ou concepts jugés pertinents qui sont comparés à une liste de mots-clés, contenant actuellement plus de 11'000 articles et concepts dans quatre langues différentes (allemand, français, italien et anglais) et relatifs à chacune des positions NOGA. Cela permet aux codeurs de sélectionner un code et de l'attribuer à l'entreprise observée.

Après cette phase, les codes attribués peuvent être mis à jour ou corrigés à tout moment, sur la base des résultats des enquêtes réalisées dans le cadre de la production statistique, des sources administratives, des registres externes, des entreprises elles-mêmes ou des informations obtenues sur Internet.

Les codes définissant l'activité économique des entreprises sont tous attribués sur la base d'informations orales ou écrites fournies par les entreprises elles-mêmes. La codification se base donc principalement sur une interprétation des textes de la part des codeurs, en la lecture, la compréhension et l'interprétation d'un texte, suivie de la définition de termes ou de concepts qui sont comparés à une liste de mots clés liés aux codes de classification.

4 Objectifs

Avec le système NOGAuto, la section URD veut construire un système d'apprentissage machine dans le but final d'automatiser l'attribution des codes d'activité économique aux entreprises enregistrés dans le SBER. Cela permettra de

- réduire au minimum l'interprétation faite par les codeurs des textes décrivant les activités économiques des entreprises,
- harmoniser et normaliser l'attribution des codes et
- minimiser le temps consacré à l'activité de codage.

NOGAuto n'est pas construit dans un seul bloc, mais, comme un oignon, en plusieurs couches. Le noyau central de l'oignon, première phase de construction qui est partie des projets pilotes de la stratégie d'innovation sur les données de l'OFS, permet de valider les codes actuellement associés aux unités déjà enregistrées dans le SBER.

La deuxième couche permettra aux codeurs d'évaluer des propositions de codes pour les activités des sociétés à codifier. L'interaction des codeurs, qui accepteront ou rejetteront les codes proposés par le système, ainsi que la saisie continue de nouveaux mots-clés proposés grâce à l'intelligence artificielle, permettront au système, avec la construction de couches supplémentaires, d'améliorer continuellement la qualité des codes proposés et d'atteindre un stade où les codes pourront, dans la plupart des cas, être attribués automatiquement, sans aucune intervention humaine.

La Figure 1 ci-dessous représente les processus du projet, dont le but est de soutenir et faciliter le codage de l'activité économique des entreprises à l'OFS.

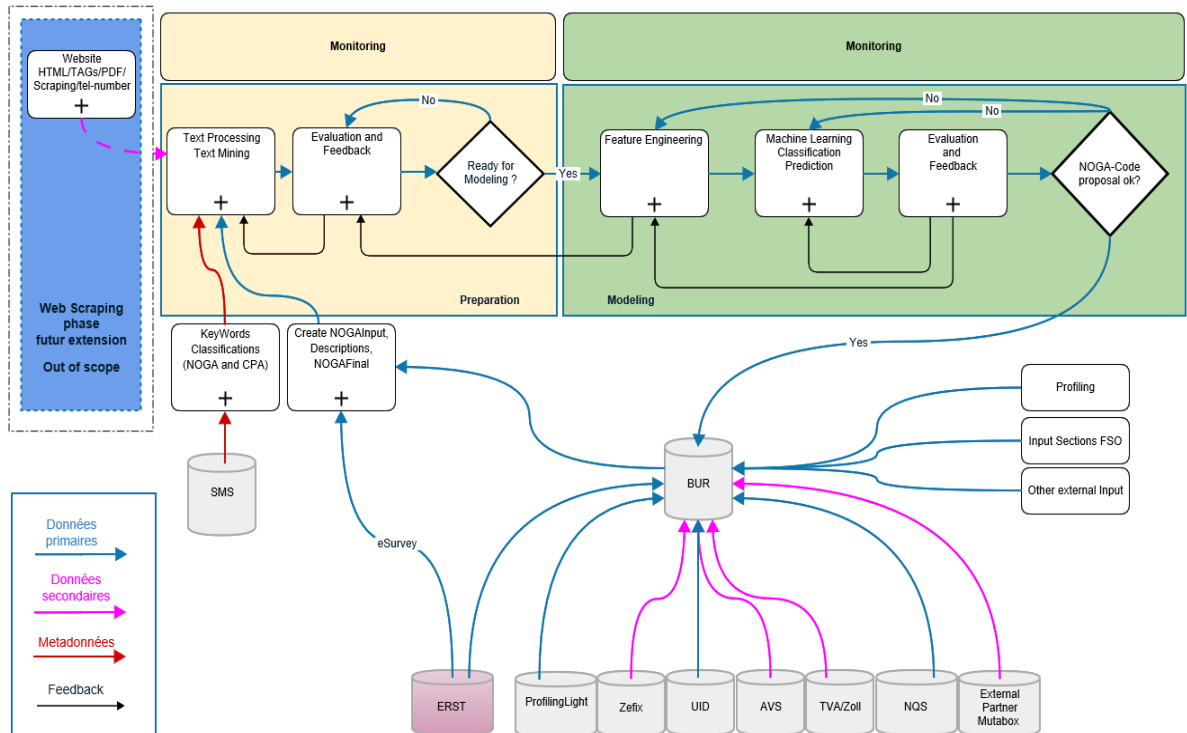


Figure 1 : Le processus du système NOGAuto avec ses éléments.

5 Résultats

5.1 Les éléments de NOGAuto et leur construction

Comme présenté dans la Figure 1 ci-dessus, le processus a été subdivisé en deux parties, la « Préparation » et le « Modeling ».

- La phase « Préparation »

Les textes décrivant les activités économiques des entreprises enregistrées dans le SBER, celles des enquêtes électroniques et les métadonnées de système central de métadonnées de l'OFS ont été importés dans le système NOGAuto et soumis au traitement automatique du langage naturel (NLP). Dans ce contexte, les textes ont été tout d'abord détectés et séparés par langue, puis standardisés (p.ex. élimination des chiffres, de la ponctuation), et nettoyés des mots vides ainsi que de ceux qui ne peuvent être directement liés à une activité économique (p.ex. articles, acronymes, prépositions).

Les résultats des opérations NLP ont ensuite été analysés et les diverses informations croisées. Une attention particulière a été accordée aux positions « complexes » de la NOGA, qui sont plus difficiles à codifier et où plus d'erreurs de codage sont commises.

Nos analyses se sont concentrées spécialement sur les activités tels que le commerce (la séparation entre le commerce de gros et de détail n'est pas toujours très claire dans les textes en provenance des entreprises) ainsi que les activités financières, immobilières et les conseils de gestion (il est souvent très difficile de clairement différencier ces activités).

La Figure 2 présente, par exemple, un extrait de l'analyse des positions de NOGA 45 « Commerce et réparation d'automobiles et de motocycles », 46 « Commerce de gros, à l'exception des automobiles et motocycles » et 47 « Commerce de détail, à l'exception des automobiles et motocycles ».



Figure 2 : Mots-clés de référence croisée NOGA 45, 46 et 47.

On peut constater que les mots clés « atelier », « carrosserie », « automobile », qui ne devraient identifier que la position 45, aient également été liés aux entreprises codifiées aux positions 46 et 47.

Ces croisements permettent également de découvrir des éléments qui n'ont pas été complètement nettoyés lors des opérations NLP (p. ex. mots dans d'autres langues, lettres isolées).

Ces images sont aussi le déclencheur de nouvelles phases de traitement du langage pour corriger les erreurs dans les mots-clés. Et cette boucle a été répétée systématiquement jusqu'à ce que les mots-clés n'aient pas atteint une qualité jugée suffisante pour lancer la phase suivante du processus.

Cette rétroaction continue à la phase précédente du processus est un élément clé de la philosophie de NOGAuto, qui est basée sur une correction et une amélioration continue des résultats des étapes précédentes.

- La phase « Modélisation »

Les unités à codifier sont reliées dans une matrice unique aux mots-clés résultant de la description de l'activité économique produite dans la phase « Préparation ». Pour affiner ultérieurement la codification, la matrice sera complétée, dans la prochaine étape du projet, par des variables supplémentaires (p.ex. adresse, emplois, chiffre d'affaires, forme juridique) en provenance du SBER. Il est ainsi prévu de relier les adresses, les emplois, le chiffre d'affaire et autres avec les descriptions des activités économiques pour permettre une ultérieure validation et affinage de la codification. Une entreprise de production de vin à avec 60 employés et un chiffre d'affaire de plusieurs millions de CHF ne pourrait, p. ex., en aucun cas être situé à St. Moritz à 1'800 m d'altitude.

A partir de cette matrice et sur la base des mots-clés et des concepts y contenus, un modèle de prédiction (« predictive analytics ») attribuera les différents codes d'activité aux entreprises. La recherche du meilleur modèle et des algorithmes appropriés pour NOGAuto est encore en cours. Des tests avec différents modèles ont eu lieu. Le modèle choisi sera lui aussi enrichi en permanence d'éléments issus des descriptions des activités des nouvelles sociétés à codifier ou d'informations complémentaires en provenance du SBER.

Une fois de plus dans ce stade aussi, il est essentiel d'obtenir un retour d'information sur la phase précédente du processus afin de corriger et d'améliorer en permanence le modèle.

La dernière étape de cette phase est l'évaluation des prédictions que le système propose. Dans cette phase de test les codes proposés sont comparés aux codes attribués manuellement par les codeurs. Toute erreur est analysée et les mots-clés actualisés.

5.2 Les premiers résultats

Les travaux dans cette première phase du projet se sont concentrés principalement (1) sur la préparation et l'amélioration continue des activités réalisées dans le cadre de la phase « Préparation », il s'est agi principalement de travaux dans le domaine du NLP (p. ex. standardisation et nettoyage des textes décrivant les activités économiques d'entreprises déjà codifiés dans le SBER) ainsi que (2) sur la recherche des meilleures méthodes de prédiction (p.ex. forêts aléatoires, réseaux de neurones) pour la modélisation des données.

Plusieurs jeux de tests ont été lancés et pour chaque nouvel essai l'input en provenance de la phase « Préparation » a été actualisé et amélioré en corrigeant les mots clés qui avaient porté sur des fausses codifications dans les essais précédents. Ces tests ont toujours été effectués avec un set de données limité (env. 1'500 enregistrements), en se limitant sur des codes NOGA à 2 positions (« digits ») et en intégrant des positions de la nomenclature à l'origine de difficultés de codification dû aux descriptions pas toujours très claires et complets de l'activité économique.

L'évolution des résultats obtenus est prometteuse. Si dans le cadre des premiers essais les prédictions du système se situaient autour de 50%, dans le dernier jeu de test – basées exclusivement sur les descriptions des activités, donc sans l'aide de variables supplémentaires – les codes corrects proposés par NOGAuto ont atteint 89%.

Bien que ces chiffres soient encore loin des critères de succès (99% de propositions de codes correctes) définies au début du projet, il faut remarquer que les travaux effectués jusqu'à présent se sont concentrés sur les cas de codification compliqués et difficiles à résoudre. Le système, déjà à l'état actuel, est en état de proposer des codes avec un pourcentage de précision nettement plus haut pour les entreprises actives dans des domaines plus faciles à décrire et de conséquent à codifier (p.ex. restaurants, hôtels, salons de coiffure).

L'Annexe 1 présente les résultats des tests de prédiction du système effectués selon les méthodes forêt aléatoire et réseau de neurones.

6 Ressources

Le projet a été réalisé exclusivement avec des ressources internes OFS en provenance des sections URD et SR de l'OFS. L'équipe a été composée par les personnes suivantes :

- Chételat Michel, SR, modélisation du système de prédiction, analyse des résultats ; Infrastructure IT ;
- Duc Sfez Cindia, URD, préparation et analyse de mots clés, analyse des résultats ;
- Joyon Christophe, URD, modélisation du système de prédiction ; analyse des résultats ;
- Macchi Claude, URD, préparation et analyse de mots clés, analyse des résultats, responsable du projet ;
- Daniele Marx, URD, modélisation du système de prédiction, analyse des résultats ;
- Ray Amadio, stagiaire URD, préparation et analyse de mots clés, analyse des résultats.

Le support expert du projet a été assuré par le Prof. Dr. Diego Kuonen, Statoo Consulting.

7 Planification et organisation

Les 12 mois à disposition pour le projet ont permis de faire une analyse de la problématique, de définir les processus, de prendre connaissance des outils informatiques à disposition ainsi que de faire des premiers essais de codification de l'activité économique des entreprises basés sur un système d'apprentissage automatique.

Ces travaux se sont concentrés principalement autour des codes NOGA les plus problématiques, ceux (1) qui contiennent des activités économiques complexes et difficiles à décrire ainsi que (2) les activités difficiles à isoler, pour leur similitude à d'autres activités (p.ex. commerce de gros et intermédiaires du commerce).

Deux éléments ont eu un effet freinant sur les travaux :

- Les ressources :
Les travaux prioritaires dans la section URD (préparation des données administratives pour les divisions de production, les difficultés informatiques dans le cadre de l'enquête ERST, la gestion de la NOGA ainsi que, dès février 2019, le projet de révision de la NACE) ont empêché un engagement majeur de ressources dans le projet NOGAuto.
- L'infrastructure :
La configuration de l'infrastructure avait été dimensionnée de façon trop faible au début du projet. Le corpus des mots-clés (créé par la description de l'activité) est très volumineux. Une première augmentation avait donné la possibilité de traiter les premiers tests avec 1'500 enregistrements. Par contre la totalité des codes NOGA ne peut pas être exécuté sur cette infrastructure. Suite au manque des ressources les pistes pour réduire les dimensions en sélectionnant que les mots-clés les plus importantes n'ont pas pu être poursuivis. La configuration de l'infrastructure n'a pas été mise au maximum, car pour les tests effectués jusqu'à maintenant la configuration a été suffisante.

La continuation du projet est étroitement reliée avec la disponibilité de ressources, soit pour la partie technique (recherche des modèles et des algorithmes appropriés) soit pour le travail de traitement des textes et de préparation et correction des mots-clés. Les ressources actuellement à disposition sont insuffisantes pour le développement d'un système tel que NOGAuto.

Une structure permanente à l'intérieur de l'OFS qui puisse garantir un suivi technique constant et une infrastructure à la hauteur est de plus un requis fondamental.

8 Conséquences/Lessons learned

- Suite à l'expérience acquise pendant ces 12 mois, on peut affirmer que la construction du système NOGAuto est absolument faisable. Les résultats atteints ont démontré que les processus ainsi que les différentes étapes de production définis sont réalistes et pourraient sans autres former la structure pour la continuation du projet.
- Bien qu'absolument faisable, la projet NOGAuto ne pourrait en aucun cas être réalisé avec les ressources actuelles. Surtout dans une première phase, dans laquelle (1) la préparation d'une première liste consolidée de mots-clés pour la totalité des positions NOGA ainsi que (2) un choix d'à implémenter dans le cadre de NOGAuto soit fait, une équipe dédiée à ce projet à raison de 3 jours par semaine pour une durée de 3 mois doit pouvoir être libérée.
- Les résultats atteints jusqu'à maintenant prouvent la valeur ajoutée qu'un système NOGAuto aurait sur le processus de codification des entreprises. Même une codification assistée dans la moitié des cas contribuerait à réduire l'interprétation faite par les codeurs, à harmoniser les l'attribution des codes et à réduire le temps de codification des entreprises.

DI projet pilote NOGAuto REG

- Un élément fondamental de tout projet d'innovation sur les données, et ceci a été confirmé aussi dans le cadre de NOGAuto, est la communication. Très souvent, les mots « automatisation » et « apprentissage automatique » sont perçus par beaucoup de gens comme synonyme de « perte de responsabilité dans le travail » ou directement « perte du travail », ce qui engendre une forte opposition à tout projet de ce genre. Il serait important de promouvoir une stratégie de communication et d'information ciblée et générale dans tout l'office pour ainsi éviter des incompréhensions et freiner des projets d'innovation.
- NOGAuto n'est pas limité à la codification des activités économiques basé sur la NOGA mais un système qui peut être adapté avec peu d'efforts pour la codification de chaque objets observé (p. ex. personnes, entreprises) avec toute nomenclature statistique (p. ex. nomenclatures des professions, des maladies, des produits).

9 Annexe

9.1 Annexe 1 : Résultats des tests effectués selon les méthodes forêt aléatoire et réseau de neurones

De suite, les résultats de tests effectués selon les méthodes de forêt aléatoire et de réseau de neurones avec un jeu d'environ 1'300 enregistrements comprenant la code NOGA à 2 positions et le texte descriptif de l'activité annoncée. (Ces enregistrements sont toujours divisés aléatoirement en un jeu d'entraînement et un jeu de test dans la proportion 2/3 – 1/3 respectivement) :

1. Forêt aléatoire

Confusion Matrix

	Reference					
Prediction	46	62	64	68	86	
46	75	1	5	2	0	
62	6	87	1	3	1	
64	3	0	75	3	3	
68	2	1	6	80	1	
86	4	1	3	2	85	

Accuracy : 0.8933 ← Taux de prédiction

2. Réseau de neurones

Taille : 5
Ecart : 0.1
Max. itérations : 100

Confusion Matrix

pred	46	62	64	68	86	
46	91	2	2	1	3	
62	9	87	3	1	3	
64	0	0	77	3	1	
68	3	0	11	82	0	
86	8	1	4	6	102	

> taux
[1] 0.878 ← Taux de prédiction

Dans les deux cas, la « Confusion Matrix » indique pour les têtes de colonne la valeur de la NOGA (2 positions) réelle et pour les lignes, la NOGA proposée.

Conclusion :

Bien que très prometteurs, les résultats ci-dessus obligent à continuer la recherche pour trouver le meilleur modèle pour NOGAuto. De nouveaux tests avec d'autres modèles sont indispensables pour trouver la confiance optimale dans la prédiction, surtout en prévision de l'intégration (1) de la totalité des positions NOGA ainsi que (2) des variables supplémentaires (voir chapitre 5.1 ci-dessus) pour affiner les modèles et ainsi améliorer la prédiction.