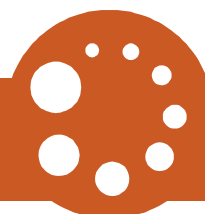




Le projet d'innovation des données «NOGAuto»

Rapport de mise en production partielle

EXPERIMENTAL STATISTICS



Neuchâtel, 2024

Éditeur: Office fédéral de la statistique (OFS)
Renseignements: info.urd@bfs.admin.ch
Rédaction: Lorenz Helbling ; Mathias Constantin ;
Daniele Marx ; Cindia Duc Sfez, OFS
Domaine: 00 Bases statistiques et généralités
Langue du texte original: Français

Concept de mise en page: Section PUB
Graphiques: Section URD/IOR
Téléchargement: www.statistique.ch
Copyright: OFS, Neuchâtel 2024
La reproduction est autorisée, sauf à des fins
commerciales, si la source est mentionnée

Table des matières

1. Compréhension du projet.....	1
1.1. Détermination des objectifs.....	1
Description du projet.....	1
Contexte	1
Objectifs du projet.....	1
Critères de succès du projet.....	2
1.2. Personnel et ressources impliquées dans le projet	3
1.3. Plan de projet.....	3
2. Compréhension des données	4
Collection des données initiales	4
Description et exploration des données	4
Vérification de la qualité des données.....	4
3. Préparation des données.....	5
3.1. Détection de la langue	6
Condition de classification	6
Algorithme avec 4 packages R + Variable REE	6
Mots récurrents.....	7
Analyse des causes originelles (« root cause analysis »)	7
Conclusion.....	10
Utilisation ultérieure dans le contexte des prédictions.....	11
3.2. Traitement automatique du langage	11
Nettoyage des données.....	11
3.3. « Feature Engineering »	12
3.4. Visualisation	13
3.5. Intégration des données	14
4. Modélisation.....	14
4.1. Sélection de la méthode de modélisation	15
Gradient Boosting Machine (GBM).....	15
Hyperparamètres	15
Indicateurs de performance	15
4.2. Générer le plan d'entraînement et de test	17
4.3. Construction du modèle	17
Validation croisée et hyperparamètres optimaux	19
Exemple de résultats	20
4.4. Evaluation et validation du modèle.....	21
4.5. Modèle en allemand et italien	24
5. Révision.....	26
5.1. Processus de révision	26
6. Déploiement.....	27
6.1. Déploiement partiel	27
6.2. Déploiement futur.....	27

7. Prochaines étapes.....	31
8. Révision du projet	32
9. Annexes.....	37
Annexe N.1: Evaluation du prototype	37
Annexe N.2: Vorgehen für einen möglichen Transfer in die Produktion.....	38
Annexe N.3: Description des variables.....	47
Annexe N.4: Exploration des données	51
Annexe N.5: R packages pour la détection de la langue.....	56
Annexe N.6: Présentation graphique mosaïque.....	56
Annexe N.7: Présentation « word embedding visualisation »	57

Résumé exécutif

Le projet NOGAuto s'inscrit dans la stratégie d'innovation sur les données de l'Office fédéral de la statistique (OFS) et fait ainsi partie des principaux projets pilotes retenus dans le but d'améliorer la production statistique publique. Le programme NOGAuto a pour vocation de prédire la classification de l'activité économique des établissements et des entreprises par l'utilisation de méthodes d'apprentissage automatique. Cette classification automatisée servira de soutien aux personnes qui codent la Nomenclature Générale de l'Activité économique (NOGA) à l'OFS. Comme de nombreuses données nécessaires à l'apprentissage automatique sont disponibles au sein de l'OFS (données d'enquêtes, description d'activité au registre du commerce, notes explicatives de la classification, etc.), ce projet s'est naturellement focalisé sur l'utilisation de ces dernières.

Dans un premier temps, les données disponibles dans le Registre suisse des Entreprises et Établissements (REE) ont été sélectionnées puis préparées afin de les rendre utilisables pour la suite du processus. Parmi ces données se trouvent les descriptions d'activités économiques qui sont la principale source pour attribuer un code NOGA aux établissements et aux entreprises. Une technique de « natural language processing » a été appliquée à ces descriptions pour extraire le maximum d'informations utiles à l'algorithme de prédiction des codes NOGA. Enfin, ces données ont été utilisées pour former des modèles (un par langue) à l'aide de techniques d'apprentissage automatique, qui ont ensuite été testés sur un échantillon aléatoire afin d'évaluer la qualité de la prédiction.

Sur la base de ces modèles, différents outils ont été développés pour être transférés en production au sein de l'OFS. Un premier outil de contrôle de listes est déjà mis en place pour assurer une meilleure qualité dans les codes NOGA déjà présents dans les différents registres. Un second produit devrait suivre sous la forme d'une application qui peut être utilisée comme soutien lors du codage de la NOGA sur la base d'une description de l'activité économique.

1. Compréhension du projet

1.1. Détermination des objectifs

Description du projet

La nomenclature générale des activités économiques (NOGA¹) est un élément essentiel de la production statistique à l'OFS. La qualité de la codification NOGA des entreprises enregistrées dans le registre des entreprises et établissement (REE) a un impact direct sur les résultats des statistiques structurelles, conjoncturelles et de synthèse qui concernent les entreprises. Ces statistiques dépendent d'une codification NOGA stable, contrôlée et de qualité. Dans un souci d'amélioration continue de la codification, le projet vise à développer un outil de soutien à la codification qui pourra être amélioré par un système de feedbacks. De plus, l'algorithme développé pourrait permettre de mettre en place un système de contrôle des codes NOGA déjà attribués aux entreprises du registre, permettant une éventuelle correction des codes mal attribués.

Contexte

Actuellement, l'attribution et le contrôle des codes NOGA se font manuellement, ce qui peut amener à des divergences dans l'interprétation des descriptions des activités économiques. Ces divergences se répercutent ensuite dans la codification, car il est parfois difficile de faire la distinction entre des codes similaires. Le codage est donc extrêmement dépendant du savoir-faire des collaborateurs les plus expérimentés² ce qui pourrait engendrer une baisse de qualité en cas de changements dans le personnel.

Le projet NOGAuto est développé au sein de la section Données des registres des entreprises (URD). Son développement est assuré par un « data scientist », un « business analyst » ainsi qu'un chef de projet. Les collaborateurs experts en codification, actifs au sein de la section, sont à disposition pour évaluer la qualité de la codification automatique du code NOGA.

Objectifs du projet

L'objectif est la création d'un programme permettant de prédire de façon automatique les codes NOGA des entreprises sur la base de leur description d'activités. Dans l'idéal, la prédiction devrait atteindre une qualité au moins égale à celle fournie par un codeur standard³.

Les bénéfices apportés par le projet pourraient être les suivants :

- Vérification automatique des codes NOGA des entreprises du REE ;
- Standardisation de la codification ;
- Réduction de l'importance de l'interprétation ;
- Réduction du temps consacré à la codification ;
- Amélioration de la qualité du codage.

De plus, la méthodologie appliquée pourrait être transposée à d'autres nomenclatures, par exemple celle des professions.

¹ <https://www.bfs.admin.ch/bfs/fr/home/statistiques/industrie-services/nomenclatures/noga.html>

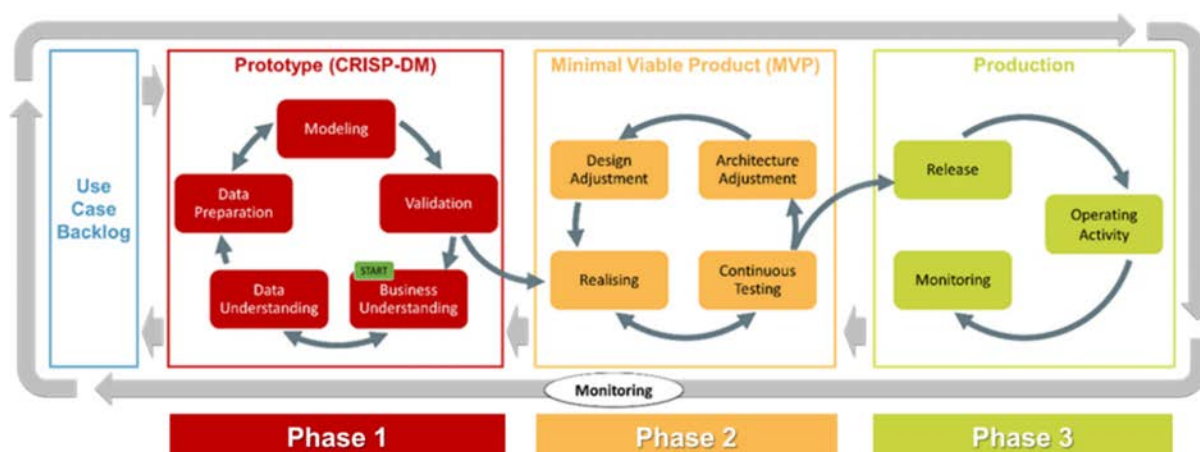
² La notion de codeur expérimenté, définie par le service des nomenclatures, correspond à une expérience de plus de 7 ans dans le domaine.

³ La notion de codeur standard, définie par le service des nomenclatures, correspond à une expérience dans le domaine entre 3 et 7 ans.

Critères de succès du projet

Pour que le projet soit réalisable, il doit remplir les conditions des trois phases⁴ présentées dans la figure suivante ainsi que les critères fixés dans la fiche « Evaluation du prototype » (annexe N.1).

Pour satisfaire aux critères du passage en production, NOGAuto doit en plus répondre aux exigences mentionnées dans le document « Vorgehen für einen möglichen Transfer in die Produktion » (annexe N.2).



Les mesures de succès nécessaires prévues initialement pour passer d'une phase à la suivante sont résumées dans la figure qui suit. Lorsque les critères listés dans la colonne Phase 1 sont remplis, on peut aborder la Phase 2, de même pour le passage de la Phase 2 à la Phase 3. La dernière colonne correspond aux critères nécessaires au maintien du projet en Phase 3.

		Phase 1	Phase 2	Phase 3
Prediction ML system-centered	Predictive accuracy	75 %	85 %	95 %
	Documentation	✓	✓	✓
	Monitoring	✓	✓	✓
Domain user-centered	Interface		✓	✓
	Performance	✓	✓	✓
	Daily use		✓	✓
	Added value for the user	✓	✓	✓

La « predictive accuracy » est une mesure globale de précision du modèle qui sera détaillée dans le chapitre 4. Ce pourcentage correspond aux nombres de prédictions correctes de l'algorithme. Les seuils ont été déterminés par le service des nomenclatures pour répondre aux exigences de la NOGA⁵.

⁴ Une description plus détaillée de ces trois phases peut être consultée sur la page <https://www.experimental.bfs.admin.ch/expstat/fr/home/innovation-science-donnees/>

⁵ Les exigences de qualité de la NOGA sont référencées dans la documentation « Detailkonzept - Revision des NOGA-Kodierungssysteme » au chapitre 5 (Qualitätsmanagement)

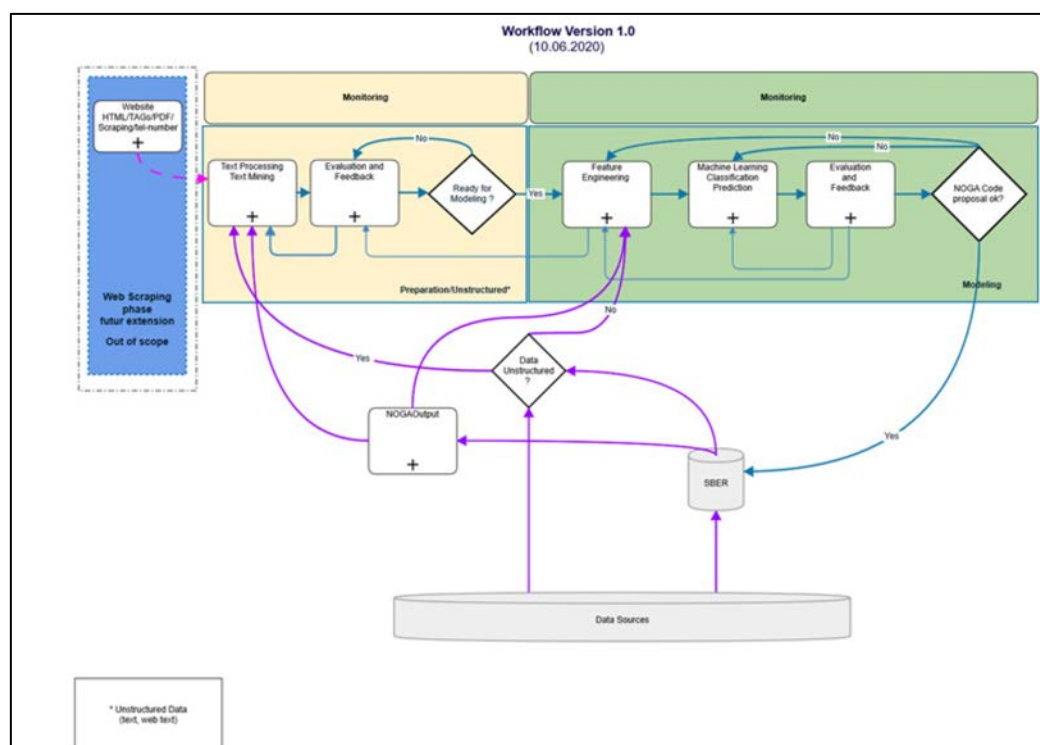
1.2. Personnel et ressources impliquées dans le projet

Personnel	• Lorenz Helbling (section URD) : Data Scientist • Mathias Constantin (section URD) : Business Analyst • 3 collaborateurs (service ERST) : Expert codage • 1 collaborateur (service NOGA) : Expert de définition du codage • Daniele Marx (section URD) : Co-chef de projet NOGAuto • Cindia Duc Sfez (cheffe de section URD) : Co-cheffe de projet NOGAuto • Athanassia Chalimourda et Daniel Assoulin (section METH): Relecture et commentaires méthodologiques • Prof. Dr. Diego Kuonen, CStat PStat CSci : Expert externe impliqué dans les premières phases du projet
Source des données	• REE
Software	• Logiciel R / Rstudio
Ressources computationnelles (hardware)	• Dans un premier lieu Atlantica, puis Renku et enfin R (sur la machine du collaborateur) lors des dernières phases du projet

1.3. Plan de projet

Le projet suit le plan détaillé dans la figure suivante.

Dans un premier temps les données du REE sont collectées, puis traitées afin de les rendre utilisables pour l'algorithme. La phase suivante correspond à l'étape de « text mining » et de « natural language processing » afin d'extraire des informations des descriptions d'activités. Enfin, des modèles d'apprentissage automatique sont entraînés afin de prédire les codes NOGA.



2. Compréhension des données

Les données à disposition proviennent du REE, qui contient des données de référence, ce qui assure leur qualité et leur fiabilité. Ces données disposent des caractéristiques suivantes :

- Une grande variété d'informations sur les entreprises ;
- Un volume conséquent concernant toutes les entreprises se trouvant sur le sol suisse ;
- Une actualité garantie par la mise à jour régulière des données du registre ;
- Une véracité assurée par les registres administratifs qui servent de source au REE.

Ces caractéristiques nous permettent d'appliquer des méthodes d'apprentissage automatique à ces données.

Collection des données initiales

Dans le but de déterminer les variables les plus intéressantes pour le projet, l'équipe des collaborateurs les plus expérimentés de l'enquête ERST (enquête auprès des nouvelles entreprises) a été consultée.

Le jeu de données contient 1'066'641 entreprises légales et 56 variables. Le jeu de données pourrait être réactualisé à tout moment en fonction des besoins.

Local from SBER									
Local_ID	Ent_ID	Name	Additional name	Registers	Universes (0,..., 7)	NOGA code	Admin statut	Stat statut	Type

Local from SBER									
Last nbr of FTJ	Nbr men	Nbr women	Street	Locality	Canton	Language	Location (est coord)	Location (nord coord)	EGID

Legal from SBER					General under the ENT_ID			Single or multi establishment	
Legal form	Registers (0,...,9)	Foundation (DD)	Foundation (MM)	Foundation (YYYY)	Turnover (attached to the ENT_ID)	Export (attached to the ENT_ID)	Import (attached to the ENT_ID)	Admin active local unit	Stat active local unit

Description et exploration des données

Le détail des 56 variables peut être trouvé dans l'annexe N3. Celles-ci sont de type numérique, texte, date ou catégorielle. Certains champs peuvent être vides pour certaines entreprises.

Les graphiques et explications relatifs à l'explorations des données se trouvent dans l'annexe N.4.

Vérification de la qualité des données

La base de données du REE est constamment mise à jour sur la base de sources diverses (dont notamment les registres administratifs), ce qui assure la qualité des observations. Cependant, afin que les variables puissent être utilisables pour le projet NOGAuto, certaines corrections ont dû être apportées :

- Remplacement des X par des 1 et des emplacements vides par des 0 dans les variables indiquant si une entreprise est active ou non ;

LEG_AVS_ACTIVE	LEG_AVS_DESCR	LEG_TVA_ACTIVE
X		
X		X
X		X
X		X

LEG_AVS_ACTIVE	LEG_AVS_DESCR	LEG_TVA_ACTIVE
1		0
0		0
0		0
1		1
1		1
1		1

- Découpage des dates de créations des entreprises en trois colonnes (jour, mois, année) ;

LEG_BIRTHDATE	day	month	year
30DEC2019:00:00:00.0000000	30	12	2019
06JUL2020:00:00:00.0000000	06	07	2020
10JUN2016:00:00:00.0000000	10	06	2016
23JAN2019:00:00:00.0000000	23	01	2019
30JAN2019:00:00:00.0000000	30	01	2019
15OCT2019:00:00:00.0000000	15	10	2019
01DEC2017:00:00:00.0000000	01	12	2017

- Suppression des observations non-conformes (par exemples valeurs manquantes) ou sans description d'activité (~125'500 observations) ;
- Changement du format de certaines variables.

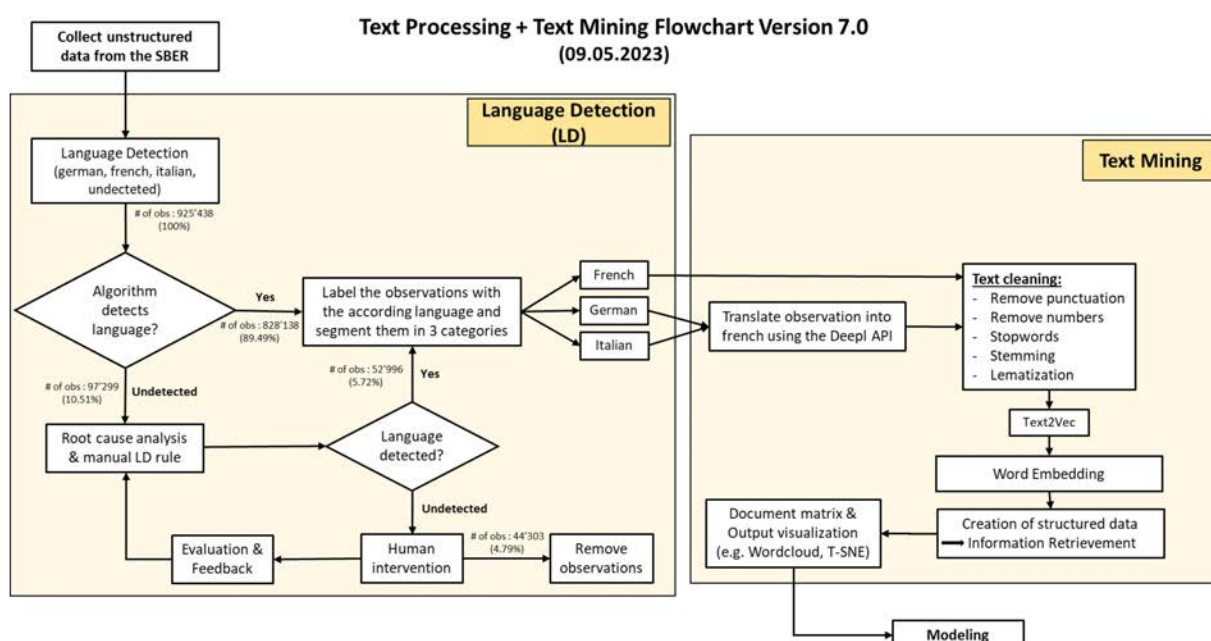
Au total 128'295 observations ont été retirées du jeu de données et certaines variables ont été renommées.

3. Préparation des données

Ce chapitre décrit la préparation des données pour leur utilisation dans la modélisation. La variable principale qui est utilisée est le descriptif de l'activité d'une entreprise. Celle-ci est la variable la plus utile aux collaborateurs pour déterminer le code NOGA d'une entreprise. A l'avenir, il pourrait être envisagé d'utiliser les autres variables disponibles (rendues utilisables pour des techniques d'apprentissage automatique par les étapes décrites au chapitre 2) comme variables secondaires. Pour le moment le projet se concentre sur la seule utilisation du descriptif d'activités.

Les descriptifs d'activités sont des données non structurées (DNS) qui sont transformées en données structurées (DS) afin de les rendre utilisables par la machine. C'est en utilisant une méthode de « text processing » que le descriptif d'activité est transformé en vecteur de texte (DS). Cette étape est décrite plus loin.

Le schéma suivant décrit les étapes pour transformer les DNS en DS.



3.1. Détection de la langue

L'algorithme de détection de la langue a pour premier but de subdiviser le jeu de données en trois en fonction de la langue de la description d'activités. Dans le REE, aucune variable n'identifie la langue de la description d'activités. La seule variable de langue à disposition ne correspond pas nécessairement à la langue dans laquelle la description d'activités est écrite. Afin qu'un modèle d'apprentissage automatique puisse prédire correctement un code NOGA sur la base de ce descriptif, il est essentiel de faire une distinction en fonction de la langue de la description. Afin de détecter automatiquement la langue, un algorithme a été élaboré. Celui-ci permet de classer les observations dans les trois langues nationales principales.

L'algorithme sert dans un premier temps à détecter la langue des descriptifs d'activités issus des données du REE afin d'entraîner des modèles spécifiques à chaque langue. Il est utilisé dans un second temps pour détecter la langue de toute description insérée dans NOGAuto afin de prédire le code NOGA associé à cette description.

L'algorithme utilise quatre packages linguistiques de R (franc, textcat, CLD2 et CLD3) pour détecter la langue. Les informations relatives à ces packages se trouvent dans l'annexe N.5.

Condition de classification

Chaque package détecte une langue. Si cette langue n'est ni l'allemand, ni le français, ni l'italien, il a été décidé d'inscrire « else » comme résultat. Si le package n'arrive pas à identifier de langue, le champ est rempli avec la mention « none ».

Grâce aux différents packages, des conditions de classifications ont été développées de façon successive.

Si au moins trois packages détectent la même langue, l'algorithme choisit cette dernière comme langue détectée. Il ne retourne pas de résultats si les conditions ne sont pas vérifiées.

1. franc=textcat=CLD2 → Langue=franc ;
➤ *Si la langue détectée par les packages franc, textcat et CLD2 est la même, alors la langue détectée sera celle du package franc.*
2. franc=CLD2=CLD3 → Langue=franc ;
3. textcat=CLD2=CLD3 → Langue=textcat.

Cette méthode permet d'attribuer une langue pour une majorité des cas. L'algorithme détecte la langue de 79.72 % des descriptifs d'activités et les 20.28 % restants n'ont pas pu être classifiés.

	Language detected (in %)	Language still not detected (in %)
Language detection algorithm	79.72	20.28

Pour améliorer la proportion d'observations dont la langue a été détectée, des techniques supplémentaires ont été mises en place.

Algorithme avec 4 packages R + Variable REE

Un nouveau critère utilisant la variable « langue » du REE est ajouté à l'algorithme pour permettre de distinguer la langue d'une partie des observations restantes. La variable issue du REE est celle évoquée au chapitre précédent.

Si la première étape n'a pas permis de déterminer la langue de la description d'activité et si la langue de la variable REE est identique à celle identifiée par deux packages, alors l'algorithme la choisit comme langue détectée.

1. REE=Textcat=CLD2 → Langue=REE;
➤ Si la langue détectée par les packages textcat et CLD3 est en accord avec la variable REE alors la langue détectée sera celle de la variable REE.
2. REE=Textcat=CLD3 → Langue=REE;
3. REE=Textcat=Franc → Langue=REE;
4. REE=CLD2=CLD3 → Langue=REE;
5. REE=CLD2=Franc → Langue=REE ;
6. REE=CLD3=Franc → Langue=REE.

Cette méthode permet d'augmenter de 10% le nombre d'observations où une langue a pu être attribuée.

	Language detected (in %)	Language undetected (in %)
Language detection algorithm + REE variable	89.49	10.51

Mots récurrents

A ce stade environ 10% des descriptions n'ont pas pu être classifiées selon leur langue. Une liste des mots les plus utilisés dans ces descriptions a été créée. Parmi ces mots, certains sont facilement identifiables par une personne alors que l'ordinateur a été incapable de les identifier. Les premiers mots récurrents identifiables par l'homme ont été sélectionnés pour établir une liste. Si un des mots dans le tableau suivant apparaît, la langue est automatiquement attribuée. L'allocation de la langue à chaque mot a été effectuée avec l'approbation de quatre collaborateurs des services NOGA et ERST.

Mot	Langue	Mot	Langue	Mot	Langue
Médecin	Français	Revier	Allemand	Transporte	Allemand
Beratung	Allemand	Architekturbüro	Allemand	Photographe	Français
Advokatur	Allemand	Coiffeursalon	Allemand	Avocate	Français
Fotografie	Allemand	Musiker	Allemand	Menuiserie	Français
Kosmetik	Allemand	Produktionsstätte	Allemand	Grafik	Allemand
Medizinalperson	Allemand	Praxis	Allemand		

Avec cette technique, 2,22% d'observations supplémentaires ont pu se voir attribuer une langue.

Analyse des causes originelles (« root cause analysis »)

L'analyse des causes originelles permet d'ajouter des combinaisons de variables qui peuvent aider à détecter une langue pour les descriptifs restants. Les techniques utilisées sont les règles d'associations, le graphique mosaïque ainsi que l'analyse de correspondance.

Règles d'associations

Grâce au package R « arules », spécialisé dans les règles d'associations, six règles supplémentaires ont pu être ajoutées à l'algorithme. Les résultats issus de l'utilisation du package sont synthétisés dans le tableau suivant.

Règle	Textcat	CLD3	CLD2	Franc	REE	Text	Résultat	«Support»	Nombre de Cas	Confiance	«Lift»
R1	-	-	Else	allemand	allemand	«Physiotherapie»	allemand	0.017	1376	1	4.18
R2	allemand	Else	None	Else	allemand	-	allemand	0.0185	1433	1	1.4
R3	Else	Else	None	italien	italien	-	italien	0.0167	1268	0.9906	1.18
R4	Else	-	-	allemand	allemand	-	allemand	0.201	15040	0.961	1.15
R5	allemand	-	-	-	-	«Verkauf»	allemand	0.104	622	1	9.43
R6	Else	-	-	Else	allemand	«Hebamme»	allemand	0.0101	595	1	5.31

Exemple d'interprétation du tableau pour la règle R1:

1. Règle R1: Si Cld2=Else et Franc=REE=allemand et que le text est « Physiothérapie » alors Langue=allemand.

Seules les règles avec un indicateur de confiance⁶ proche de 1 (valeur maximale) ont été sélectionnées.

Après l'ajout de ces règles, le pourcentage d'attribution de langue à des descriptions d'activités augmente comme suit.

	Language detected (in %)	Language undetected (in %)
After 1 st rule	91.71	8.29
After 2 nd rule	91.81	8.19
After 3 rd rule	91.94	8.06
After 4 th rule	93.57	6.43
After 5 th rule	93.64	6.35
After 6 th rule	93.70	6.30

Graphique mosaïque

Le graphique mosaïque est une autre technique qui peut être utilisée dans le contexte de la détection de la langue. Cette méthode d'analyse visuelle de graphiques permet de repérer certaines combinaisons de variables⁷. Grâce à ce type de technique, le nombre d'observations non classifiées dans une des trois langues a pu être réduit. Le tableau ci-dessous montre les combinaisons issues de l'utilisation du package R « vcd » avec la fonction « mosaic ». Chaque ligne correspond à un graphique mosaïque. Un exemple de graphique se trouve dans l'annexe 6.

Mosaic	Textcat	CLD3	CLD2	Franc	REE	Result	Pearson Res	# Cases
Mosaic1	Else	None		French	French	French	99	5'275
Mosaic2		Else	None	French	French	French	91	3'069
Mosaic3		Else	Else	French	French	French	92	1'220
Mosaic4		None	None	Italian	Italian	Italian	93	983

Exemple pour la première combinaison:

- Textcat=Else & Cld3=None & Franc=REE= Français → Langue=Français
 - Si Textcat=Else et Cld3=None et Franc ainsi que REE détectent la langue française alors le résultat sera « Français ».

Après avoir ajouté ces règles à l'algorithme de détection de la langue, nous observons une augmentations du nombre d'observations pour lesquelles une langue a pu être attribuée.

	Language detected (in %)	Language undetected (in %)
Mosaic 1	94.27	5.73
Mosaic 2	94.60	5.40
Mosaic 3	94.73	5.27
Mosaic 4	94.84	5.16

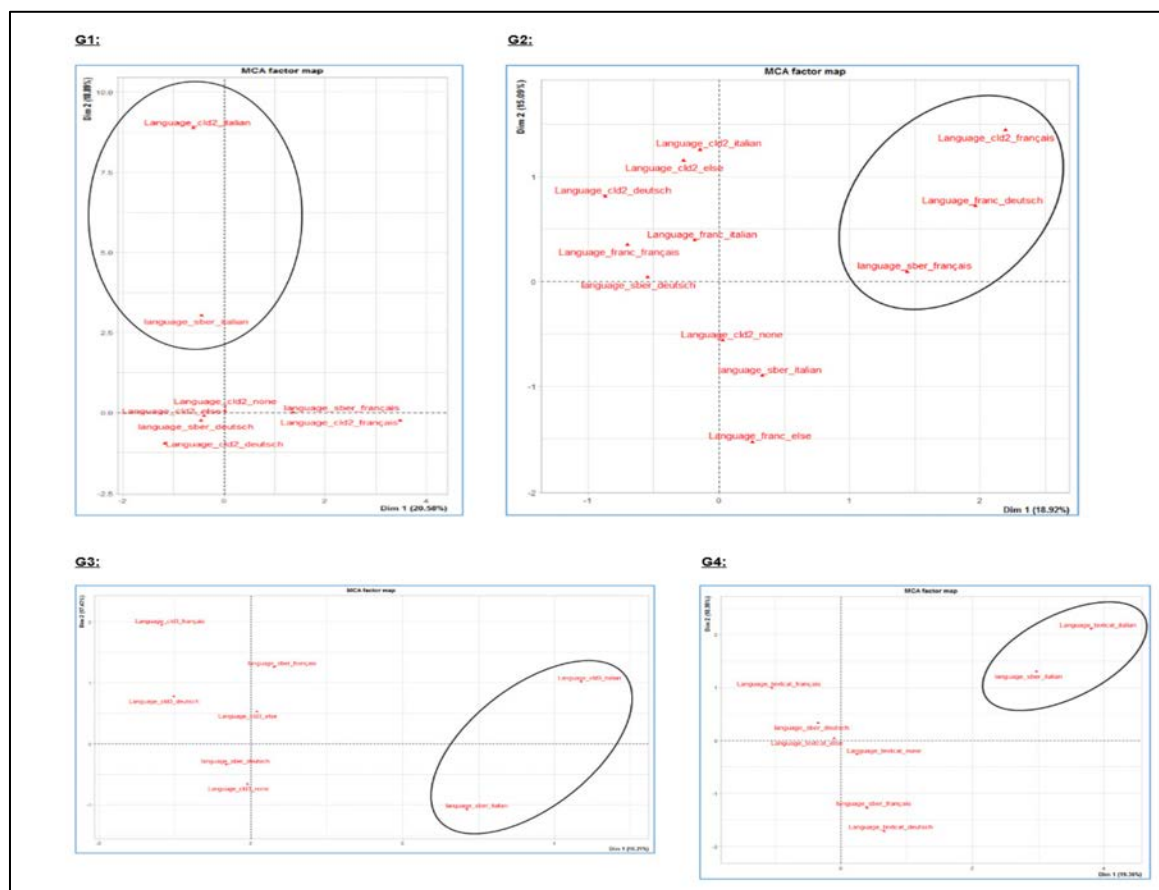
⁶ <https://mhahsler.github.io/arules/docs/measures#confidence>

⁷ Pour plus de précisions sur l'interprétation des graphiques mosaïques, on peut consulter l'article suivant : [Mosaic Displays for Multi-Way Contingency Tables \(datavis.ca\)](https://www.datavis.ca/mosaic/)

Analyse de correspondance

L'analyse de correspondance est également une technique d'interprétation visuelle de graphiques. Pour appliquer cette méthode, c'est le package R « FactoMineR »⁸ qui a été utilisé. On recherche des groupes de points similaires dans les graphiques avec la plus grande somme de variance possible.

Le tableau ci-dessous expose les 4 graphiques extraits avec (avec la fonction « mca ») ainsi que les regroupements de points considérés.



Les règles qui ont été déduites de l'interprétation de ces graphiques sont les suivantes (où la somme de la variance totale est définie par les deux premières dimensions) :

Plot	Combinations	Result	Somme de la variance	# of Cases
G1	REE_it & CLD2_it	Italian	39.47%	125
G2	REE_fr & CLD2_fr & franc_all	French	34.01%	829
G3	REE_it & CLD3_it	Italian	35.64%	169
G4	Textcat_it & REE_it	Italian	37.75%	310

- Dans le graphique G1, il est possible de voir que la variable REE et CLD2 sont dans la même zone quand il est question de détecter la langue italienne avec une variance totale de 39.47%.

Grâce à cette technique, le pourcentage d'observations dont la langue a pu être détectée est monté de 0.15%.

⁸ <http://factominer.free.fr/factomethods/analyse-des-correspondances-multiples.html>

	Language detected (in %)	Language undetected (in %)
Correspondence analysis 1 (G1)	95.06	4.94
Correspondence analysis 2 (G2)	95.16	4.84
Correspondence analysis 3 (G3)	95.18	4.82
Correspondence analysis 4 (G4)	95.21	4.79

Conclusion

Après avoir ajouté les règles déterminées par ces analyses de causes originelles, la quantité de descriptions dont la langue a pu être détectée est montée à 95,21%. Tout au long du développement de l'algorithme, un contrôle visuel sur 200 unités aléatoires a été effectué afin de vérifier si la langue détectée correspond à la langue de la description. Ce contrôle n'a identifié aucune erreur d'attribution de la langue par l'algorithme.

	Language detected	Language undetected
Number of enterprises	881'134	44'303

Après application de l'algorithme sur notre jeu de données, la répartition par langue est la suivante.

- 625'909 entreprises avec une description d'activité en allemand;
- 208'999 entreprises avec une description d'activité en français;
- 46'226 entreprises avec une description d'activité en italien.

La langue de 44'303 observations n'a pas pu être détectée à l'aide de l'algorithme. Il a été décidé de vérifier si la variable langue présente dans le REE peut être utilisée pour classer les observations restantes. La répartition selon cette variable est la suivante.

- Allemand: 31'690;
- Français: 10'610;
- Italien: 2'003.

Cette solution n'est pas applicable en partie à cause de la présence de descriptions rédigées en anglais. Sur toutes les observations restantes, un test aléatoire sur 200 unités par langue a été effectué et nous donne le résultat suivant.

Dataset	Mauvaise classification (en %)	Mauvaise classification sans les descriptions en anglais (en %)
German	21/200 = 10.50%	7/200 = 3.5%
French	24/200 = 12%	8/200 = 4%
Italian	44/200 = 22%	20/200 = 10%

Les résultats peu concluants ne nous permettent pas d'utiliser la langue issue du REE seule pour attribuer une langue aux observations restantes.

Cet algorithme n'est pas la partie principale du projet NOGAuto, mais un produit additionnel développé pour les besoins du projet. La méthode peut tout à fait faire l'objet d'améliorations afin d'augmenter la quantité de descriptions dont la langue a pu être détectée.

Utilisation ultérieure dans le contexte des prédictions

L'algorithme de détection de la langue est également utilisé dans le cadre des prédictions effectuées par NOGAuto. Il permet de détecter la langue d'une activité introduite dans le système. La variable REE étant inexistante pour ce type d'utilisation, seule la première condition de classification ainsi que les mots récurrents sont utilisés dans ce contexte. Cependant le collaborateur devrait pouvoir sélectionner manuellement une langue si NOGAuto ne parvient pas à la détecter correctement. Ce point est détaillé dans le chapitre consacré au déploiement futur (chapitre 6).

3.2. Traitement automatique du langage

Appelée en anglais « text mining », cette étape consiste en une extraction d'informations issues des descriptifs d'activités permettant de détecter les tendances en fonction des codes NOGA.

Dans un premier temps, les données sont nettoyées. Les différents mots sont ensuite transformés en vecteurs numériques.

La méthodologie a été élaborée sur la base du jeu de données en français.

Nettoyage des données

Les étapes de nettoyage des textes sont les suivantes :

- Transformation de tous les caractères en minuscules ;
- Suppression des mots neutres (répertoriés par les équipes ERST et NOGA) et de certains mots spécifiques (société, statuts, exploitation, ...) – La liste de mots peut être mise à jour selon les requêtes des services concernés ;
- Suppression de la ponctuation ;
- Suppression des nombres ;
- Suppression des espaces inutiles ;
- Réalisation d'un « stemming » qui permet de réduire les mots à leur « racine » en éliminant les dernières lettres (par exemple « collaborateur » devient « collab ») ;
- Remplacement des caractères spéciaux par des caractères standards (è devient e, ï devient i, ...).

Voici un exemple de descriptions d'activités avant et après ces étapes de nettoyage.

vente et commerce de marchandises, notamment de produits bureautiques; exploitation d'un local de vente et d'un site de vente en ligne et vente par correspondance; gestion et exploitation de café-restaurant, de boulangerie et de tea-room; import-export et achat-vente de divers produits alimentaires (cf. statuts pour but complet).	vent commerc marchandis produit bureaut local vent sit vent lign vent correspond gestion caf restaur boulanger te room import export achat vent produit alimentair
la société a pour but la fabrication et la distribution de produits de boulangerie, pâtisserie, confiserie, glacerie et chocolaterie, ainsi que l'exploitation d'un tea-room, service traiteur, chef à domicile, conseil et développement de concepts de bouche (pour but complet cf. statuts).	fabriqu distribu produit boulanger patisser confiser glacer chocolater te room servic traiteur chef domicil conseil developp concept bouch
achat, vente de terrains; exploitation de terres agricoles, d'un commerce de raisins, de moûts, vins, spiritueux, distillés, et de toutes autres boissons alcoolisées ou non; gestion d'immeubles ainsi que toutes activités convergentes à ses buts	achat vent terrain terr agricol commerc raisin mout vin spiritui distill boisson alcoolis gestion immeubl activ convergent but
La société a pour but l'exploitation d'une entreprise de menuiserie-charpenterie. La société peut notamment effectuer la conduite ou la surveillance de travaux, réaliser elle-même ou faire réaliser par des tiers des travaux de construction ou de rénovation de biens immobiliers. Elle peut également e	entrepris menuiser charpenter effectu conduit surveil travail realis realis tier travail construct renov bien immobili egal
Exploitation d'une entreprise de menuiserie et charpente.	entrepris menuiser charpent

3.3. « Feature Engineering »

Les mots des différentes descriptions ont été rassemblés pour créer un vocabulaire propre aux descriptifs d'activités. Certains de ces mots ont été écartés car ils n'apparaissent pas plus de cinq fois dans l'intégralité du jeu de données.

Les mots sont ensuite transformés en vecteur numériques à l'aide de packages R. Les vecteurs sont générés sur la base du jeu de données et sont construits de telle façon que les mots similaires se voient attribuer une valeur proche l'une de l'autre. Cette étape est connue en français sous le terme de plongement lexical (en anglais « Word Embedding »).

A titre d'exemple le package « Word2Vec » permet de générer un vecteur pour chaque mot en utilisant des techniques de réseaux de neurones. On peut ensuite déterminer pour chaque paire de mots leur facteur de similarité. Pour illustrer cette démarche, les méthodes « Continuous Bag of Words » (CBOW) et « Skip Gram » ont été appliquées aux mots « développement » et « chantier ». Les deux méthodes nous permettent de trouver les autres mots issus du dictionnaire spécifique au projet créé plus haut dont le facteur de similarité avec « développement » ou « chantier » est le plus proche possible de 1 (valeur maximale). Les tableaux suivants illustrent ces résultats.

Méthode CBOW :

\$developp	term1	term2	similarity	rank
1	developp	recherch	0.8709304	1
2	developp	industrialis	0.8631757	2
3	developp	implant	0.8485250	3
4	developp	nouvel	0.8240800	4
5	developp	method	0.8162531	5
6	developp	processus	0.8155367	6
7	developp	progiciel	0.8084284	7
8	developp	innov	0.8078639	8
9	developp	implement	0.7746829	9
10	developp	business	0.7674263	10

\$chanti	term1	term2	similarity	rank
1	chanti	travail	0.8081709	1
2	chanti	surveil	0.7909160	2
3	chanti	bat	0.7876418	3
4	chanti	coordin	0.7844188	4
5	chanti	acrobat	0.7692313	5
6	chanti	thermograph	0.7684194	6
7	chanti	mesur	0.7679777	7
8	chanti	planif	0.7611527	8
9	chanti	entretien	0.7580220	9
10	chanti	expertis	0.7393718	10

Méthode Skip-Gram :

\$developp	term1	term2	similarity	rank
1	developp	innov	0.8921235	1
2	developp	integr	0.8809921	2
3	developp	technolog	0.8795009	3
4	developp	solut	0.8785451	4
5	developp	concevoir	0.8775232	5
6	developp	recherch	0.8697035	6
7	developp	industrialis	0.8625280	7
8	developp	reproduir	0.8569233	8
9	developp	implant	0.8551753	9
10	developp	semiconducteur	0.8531486	10

\$chanti	term1	term2	similarity	rank
1	chanti	bat	0.8855168	1
2	chanti	incend	0.8718788	2
3	chanti	vitrag	0.8701755	3
4	chanti	enduit	0.8600124	4
5	chanti	utilitair	0.8588041	5
6	chanti	ascenseur	0.8587046	6
7	chanti	traval	0.8490586	7
8	chanti	renov	0.8481492	8
9	chanti	construct	0.8453675	9
10	chanti	exterieur	0.8443334	10

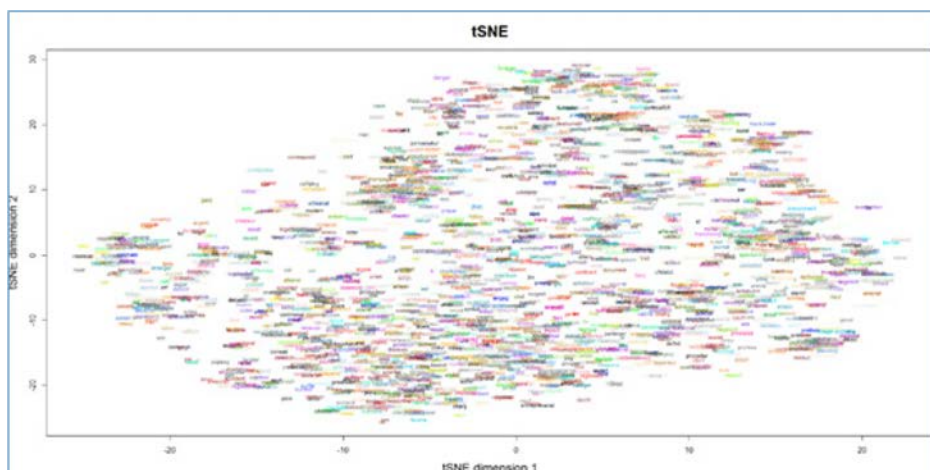
Dans le contexte de NOGAuto, c'est le package « Text2Vec » qui est utilisé car il permet d'associer un vecteur au texte complet. Notre jeu de données en français contient 7'141 mots qui apparaissent dans les descriptions d'activités. Dans un premier temps, le package « Text2Vec » associe à chaque mot un vecteur de la dimension désirée (1'000 dans la première phase de développement). Le package détermine sur la base de notre jeu de données si les mots ont des significations semblables ou sont issus d'un contexte similaire et leur attribue, le cas échéant, des valeurs proches l'une de l'autre. Ceci permet à la machine de mieux comprendre les relations entre les mots et améliore la performance de l'algorithme d'apprentissage automatique. Le tableau ci-dessous illustre les treize premières dimensions pour les premiers mots de la liste.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
accru	-0.08867645	0.084150699	-0.244486065	-0.289365349	-0.2290673313	-0.547957343	-0.274741816	-0.5796846414	0.5299315232	-0.0091893670	-0.246007422	-0.115860037	-0.674998459
acid	-0.43251243	-0.750237724	-0.379317391	0.706470043	0.1103143657	-0.202459695	0.308386763	0.2596851707	-0.1775962001	-0.2709142176	0.074740753	0.089537237	-0.442470183
activi	0.57526031	-0.064844107	0.320769358	-0.557466706	0.5335911951	0.049349228	-0.476825000	0.1221424498	-0.4789526614	0.2356465418	-0.234283000	0.411410465	0.895551468
actric	-0.18175933	0.176478511	0.306696514	0.515022487	0.0175651498	-0.241464775	0.048359709	0.2404724014	-0.8150094691	0.0128850332	0.207768268	0.261413968	0.398636374
actu	-0.57496487	-0.309921134	-0.248109988	-0.378388084	0.0121711661	0.294197635	0.106391409	-0.2909414109	-0.2313336895	0.0840720598	-0.067621169	-0.633320430	0.601281145
adjuv	-0.55485675	0.003681983	0.394080671	0.203467124	-0.2993191391	0.081009113	-0.011867569	0.1669176122	-0.5025888509	0.0561634141	-0.223473024	-0.175325107	-0.067941247
aerat	-0.14293069	-0.204981753	-0.009096014	0.491574278	-0.0300977660	-0.379972930	0.448153386	-0.0847059493	0.1833004206	-0.2177462297	-0.182800416	0.275558751	-0.830731612
aerograph	-0.84048575	-0.014754648	0.040063022	0.307424514	1.1767821887	-0.100497204	-0.100906954	0.2995356022	-0.2291678369	0.1449919099	0.343297555	0.373552351	0.715207834
aerol	0.02973447	0.295107998	0.347227602	0.151324949	-0.4455771170	-0.557143535	0.161393846	-0.0251655182	0.6506773558	0.3907014710	-0.444666880	0.501313233	-0.716117194
afflelou	-0.20560085	-0.187998635	-0.049046550	0.126048020	0.1639861902	-0.091951957	-0.255051045	0.5220297261	-0.5122076288	-0.0720463957	0.348516012	0.221099701	-0.114747168
affreteur	0.45507886	0.138374242	0.364429978	-0.064994910	0.4308147323	-0.022325859	0.834439690	0.4909878265	0.0084826542	-0.1069363833	-0.408806597	0.787986195	-0.251754123
afro	-0.55625169	-0.432941456	-0.032753626	0.140569992	-0.2947222704	-0.215178628	-0.003638784	-0.1473642906	0.1632598017	-0.0240708304	-0.637715787	0.808015886	-0.033631078
aging	-0.33828278	0.137817364	-0.179436716	0.498775174	0.1260348367	-0.521606084	-0.042667329	0.6755255748	0.3026313620	-0.2652705493	0.685938049	0.355179055	-0.104785854
agress	0.24265480	-0.739649564	0.008509471	-0.470389162	0.4125146222	-0.245956990	0.207771928	0.1672677261	-0.2735129463	0.0232536418	-0.130930015	-0.074824478	-0.204122175
agritour	0.44914649	-0.158278641	0.462244929	0.199342956	0.4233937792	-0.122572207	-0.014919402	0.2097048798	0.0157091017	-0.5245232111	0.410362762	0.294646920	0.038877834
aigue	-0.08175369	0.207768329	-0.746677972	0.313863875	1.0123520631	0.122771609	-0.160983133	-0.5359273518	0.3815958655	1.0017973148	-0.441904198	-0.570723550	0.050979991
ainis	-0.12976597	-0.603266389	-0.470592457	-0.242591443	0.2878165700	0.109856573	0.392004531	0.6426916701	0.6207883082	0.3504437181	0.434047824	0.369744289	0.769072070

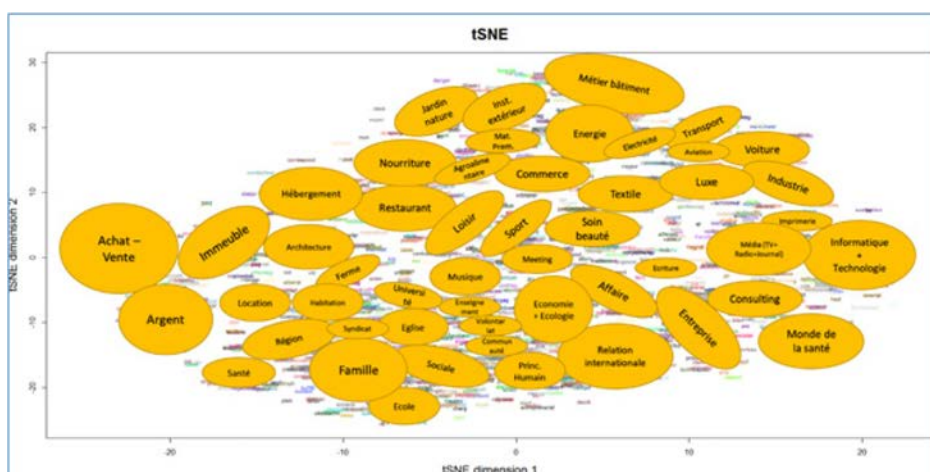
Une description d'activités peut être composée de plusieurs mots. Dans ce cas-là, la description d'activités se voit attribuer un vecteur pour la représenter. Ce vecteur correspond à la moyenne des vecteurs associés aux mots de la description.

3.4. Visualisation

Afin de contrôler visuellement que l'étape de plongement lexical a fonctionné correctement, il est possible de créer des graphiques à l'aide du package R « T-SNE ». Ce package permet de classer les mots similaires dans les mêmes zones en réduisant la dimension du vecteur qui représente chaque mot. Les détails des vues du graphique T-SNE se trouvent dans l'annexe N.6.



Cette technique de visualisation nous permet également d'établir un graphique illustrant les regroupements de mots selon une même thématique. En effectuant un contrôle visuel, chaque zone thématique peut être recouverte d'une ellipse et permet de visualiser les champs sémantiques présents dans notre jeu de données. Par exemple, dans le graphique suivant, les mots auto, carrossier, moteur, garage, ... peuvent être trouvés sous l'ellipse labellisée « voiture ».



3.5. Intégration des données

Après ces étapes de préparation, le jeu de données peut être utilisé pour la modélisation à l'aide de techniques d'apprentissage automatique. Grâce à la transformation des mots et descriptifs sous forme de vecteurs, les caractéristiques sémantiques de ceux-ci peuvent être interprétées par un ordinateur.

4. Modélisation

Ce chapitre décrit les techniques d'apprentissage automatique qui sont utilisées pour entraîner l'ordinateur à créer des modèles de prédiction. Ces modèles sont ensuite utilisés pour prédire les codes NOGA de descriptions d'activités.

La méthode principale utilisée est celle du « Gradient Boosting Machine⁹ » (GBM). Celle-ci a été sélectionnée sur les conseils de l'expert externe qui était présent lors des premières phases du projet. Elle présente des résultats concluants, notamment lors de compétitions de « machine learning » sur la plateforme Kaggle.

⁹ http://uc-r.github.io/qbm_regression

En raison de la capacité de calcul réduite, il a été décidé de réduire le nombre de dimensions des vecteurs de texte de 1'000 à 100 ce qui a permis à la machine de générer des modèles sans interruptions dans la phase d'entraînement. Ce nombre de dimensions choisi par défaut peut être augmenté et réévalué à l'avenir en fonction du développement de l'infrastructure.

Il est utile de rappeler que le développement méthodologique est basé sur le jeu de données avec des descriptions d'activités en français.

4.1. Sélection de la méthode de modélisation

Gradient Boosting Machine (GBM)

Cette méthode d'apprentissage automatique fait partie de la famille des arbres de décisions. Le GBM suit la logique des « boosting trees ».

Il s'agit d'une forme d'apprentissage dans laquelle les arbres sont générés séquentiellement. Les nouveaux arbres sont améliorés là où les arbres précédents n'ont pas été très performants.

La fonction de GBM entraîne le modèle en essayant de minimiser une fonction de perte, dite « loss function ».

Hyperparamètres

Chaque modèle peut être ajusté par le biais d'hyperparamètres afin de sélectionner le modèle optimal (celui qui minimise la fonction de perte). La liste ci-dessous énumère les hyperparamètres considérés pour la méthode « GBM » :

- **Subsample** : Ratio de sous-échantillonnage de l'échantillon d'entraînement. Par exemple, une valeur de subsample de 0.7 signifie que 70% des données d'entraînement sont utilisées pour générer un nouvel arbre qui est testé sur les 30% restants ;
- **Gamma** : Régulateur de la complexité de l'arbre, en définissant la réduction minimale de perte pour une séparation suivante d'une partie de l'arbre ;
- **Max_depth** : Paramètre qui limite la profondeur de l'arbre ;
- **Eta** (learning rate) : Paramètre de régularisation qui réduit les poids des variables à chaque étape d'amélioration ;
- **Min_child** : Nombre minimal de cas dans un nœud (plus ce dernier est haut, plus l'algorithme est conservateur).

De plus amples informations sur les rôles, la portée et les valeurs par défaut des hyperparamètres peuvent être trouvées sur la page internet dédiée au package « xgboost »¹⁰ (eXtreme Gradient Boosting).

Dans le but de trouver la meilleure combinaison d'hyperparamètres pour l'entraînement des modèles, une validation croisée (ou « cross validation ») a été effectuée. Cette étape est détaillée dans le chapitre 4.4.

Indicateurs de performance

Afin de pouvoir mesurer les performances des modèles de manière satisfaisante, l'équipe en charge du projet a sélectionné diverses mesures de qualité. Ces mesures de qualité utilisent les notions appelées en anglais « true positive », « false positive », « true negative » et « false negative ».

A titre d'exemple, dans le contexte d'une classification des observations en deux classes (« positive » et « negative »), quatre situations sont possibles.

¹⁰ <https://xgboost.readthedocs.io/en/stable/parameter.html>

		Prediction	
		Positive	Negative
True value	Positive	A. True Positive (TP)	B. False Negative (FN)
	Negative	C. False Positive (FP)	D. True Negative (TN)

- A. TP : Le modèle prédit la valeur « positive » et cette prédiction est correcte,
- B. FN : Le modèle prédit la valeur « negative » (ou « non-positive ») mais cette prédiction est incorrecte,
- C. FP : Le modèle prédit la valeur « positive » mais cette prédiction est incorrecte,
- D. TN : Le modèle prédit la valeur « negative » (ou « non-positive ») et cette prédiction est correcte.

Si on applique cette façon de raisonner au contexte des codes NOGA, on peut par exemple obtenir un tableau pour chaque code NOGA possible (ou chaque section NOGA). Les valeurs attribuées à chaque case de ce tableau, pour un code NOGA X sont interprétées de la sorte.

- A. TP : Le modèle prédit le code NOGA X et il s'agit du bon code NOGA,
- B. FN : Le modèle prédit un autre code NOGA que X et se trompe,
- C. FP : Le modèle prédit le code NOGA X et se trompe,
- D. TN : Le modèle prédit un autre code NOGA que X et cette prédiction est correcte.

Grâce à ces différentes possibilités de classification, nous pouvons définir les quatre indicateurs de performances qui ont été sélectionnés pour évaluer la qualité des modèles développés (dans les définitions suivantes, le symboles # est utilisé pour signifier le nombre d'occurrences d'un cas de figure, par exemple #TN correspond au nombre total de TN). La « predictive accuracy » peut être évaluée sur n'importe quel jeu de prédictions alors que les trois autres mesures seront utilisées pour mesurer la performance dans des classes ou des codes NOGA spécifiques.

1. Predictive accuracy : mesure de la proportion de prédictions correctes.

$$accuracy = \frac{\#Classifications\ correctes}{\#Classifications}$$

2. Precision : mesure de la proportion de True Positive (TP) sur le nombre de cas prédits par le modèle pour une classe donnée.

$$precision = \frac{\#TP}{\#TP + \#FP}$$

Concrètement, dans une catégorie de NOGA donnée, quel est le pourcentage d'unités pour lesquelles la prédiction de cette NOGA a été effectuée correctement. Une faible valeur de « precision » signifie que de nombreuses prédictions sont attribuées à cette NOGA alors qu'elles correspondent à un autre code NOGA.

3. **Recall**¹¹ : mesure de la proportion de True Positive (TP) sur le nombre de vraies valeurs¹² pour une classe donnée.

$$recall = \frac{\#TP}{\#TP + \#FN}$$

Concrètement, dans une catégorie de NOGA donnée, quel est le pourcentage d'unités pour lesquelles cette catégorie NOGA a été prédite par rapport au nombre réel d'observations dans cette catégorie. Une faible valeur de « recall » signifie que beaucoup de descriptions sont attribuées à tort à d'autres catégories NOGA et non à celle dont il est question.

4. **F1-score** : moyenne harmonique des deux mesures de qualité décrites ci-dessus, il synthétise donc les valeurs de « precision » et de « recall » :

$$f1 = \frac{2 * (precision * recall)}{(precision + recall)}$$

4.2. Générer le plan d'entraînement et de test

Dans des projets d'apprentissage automatique, il est courant de diviser aléatoirement le jeu de données en trois parties, « training set », « validation set » et « test set ». Dans le cadre de NOGAuto, pour entraîner et évaluer la performance des modèles, le ratio retenu est de 80%-20%. (80% de données pour générer le « training set » et 20% pour le « test set »). Le « training set » permet d'entraîner les modèles à l'aide d'une partie des données. Le modèle est ensuite testé sur la partie restante. Le « validation set », qui est inclus dans le training set, est quant à lui utilisé pour la validation croisée, qui permet de déterminer les hyperparamètres optimaux.

Dans le contexte de la modélisation en français, nous disposons d'un jeu de données composé de 209'000 observations. Pour des raisons d'infrastructure, il n'est malheureusement pas possible d'utiliser l'intégralité de ces données pour entraîner les modèles. De ce fait, afin de rendre la répartition au sein des classes un peu plus équilibrée dans le jeu de données réduit et afin de couvrir le mieux possible les petites classes, la démarche suivante a été appliquée.

- Comptage de la cardinalité des classes NOGA 4 dans le jeu de données complet ;
- Si pour une classe NOGA X, #X > 100, insertion de 100 observations pour cette classe X dans le jeu de données réduit ;
- Si pour une classe NOGA X, #X < 100, insertion de toutes les observations disponibles pour cette classe X dans le jeu de données réduit.

Après cette étape, le jeu de données réduit est constitué 32'040 observations. La taille du « training set » est de 25'632 observations avec un « test set » constitué de 6'408 observations. La distribution des 32'040 observations entre le training set et le test set a été effectuée aléatoirement.

4.3. Construction du modèle

Les premières tentatives d'entraînement de modèles de prédiction ont été faites directement sur la variable dépendante code NOGA à 4 digits. Les résultats n'ont malheureusement pas été concluants. Le nombre de catégories possible n'a pas permis un entraînement de modèles satisfaisant. Pour cette raison, une autre approche a été préférée à la première. Les modèles sont entraînés de façon successive, tout d'abord sur les sections NOGA, puis sur les deux digits, les trois digits, les quatre digits et finalement sur les six digits.

¹¹ Aussi appelée « sensitivity » dans le monde statistique

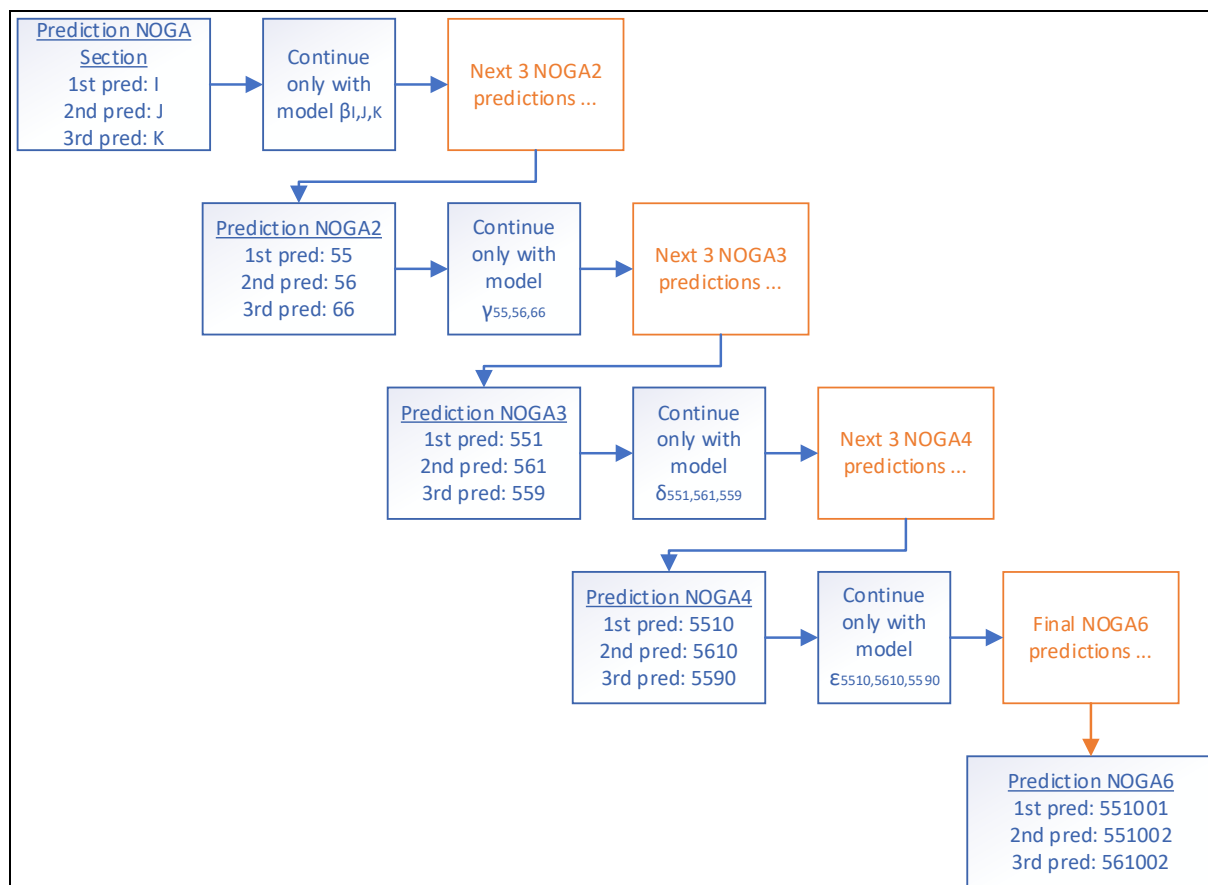
¹² C'est-à-dire, pour une classe donnée, le nombre d'observations qui appartiennent à cette classe.

L'idée est donc de procéder par étapes successives. A chaque étape, le modèle prédit les trois catégories les plus probables et les utilise pour continuer ses calculs sur la base d'un modèle entraîné uniquement pour ces trois catégories. Concrètement, l'algorithme suit les étapes ci-dessous :

1. Application du modèle α pour prédire la section NOGA de l'entreprise à partir de sa description d'activités,
2. Sélection des 3 sections NOGA (i,j,k) avec la plus haute probabilité prédite par le modèle α ,
3. Chargement d'un modèle de prédiction des codes NOGA2 $(\beta_{i,j,k})$ entraîné pour les 3 sections NOGA sélectionnées au point 2,
4. Application du modèle $\beta_{i,j,k}$ pour prédire le code NOGA2 de l'entreprise à partir de sa description d'activités,
5. Sélection des 3 codes NOGA2 (l,m,n) avec la plus haute probabilité¹³ prédite par le modèle $\beta_{i,j,k}$,
6. Chargement d'un modèle de prédiction des codes NOGA3 $(\gamma_{l,m,n})$ entraîné pour les 3 codes NOGA2 sélectionnés au point 5,
7. Application du modèle $\gamma_{l,m,n}$ pour prédire le code NOGA3 de l'entreprise à partir de sa description d'activités,
8. Sélection des 3 codes NOGA3 (o,p,q) avec la plus haute probabilité prédite par le modèle $\gamma_{l,m,n}$,
9. Chargement d'un modèle de prédiction des codes NOGA4 $(\delta_{o,p,q})$ entraîné pour les 3 codes NOGA3 sélectionnés au point 8,
10. Application du modèle $\delta_{o,p,q}$ pour prédire le code NOGA4 de l'entreprise à partir de sa description d'activités,
11. Sélection des 3 codes NOGA4 (r,s,t) avec la plus haute probabilité prédite par le modèle $\delta_{o,p,q}$,
12. Chargement d'un modèle de prédiction des codes NOGA6 $(\epsilon_{r,s,t})$ entraîné pour les 3 codes NOGA4 sélectionnés au point 8,
13. Application du modèle $\epsilon_{r,s,t}$ pour prédire le code NOGA6 de l'entreprise à partir de sa description d'activités,
14. Prédiction finale des 3 codes NOGA6 les plus probables.

Le diagramme ci-dessous illustre la procédure avec un exemple :

¹³ Il est à noter ici que cette probabilité n'est pas absolue, dans le sens où elle est conditionnée par le choix des trois sections NOGA les plus probables effectué au niveau précédent. I.e. la probabilité affichée correspond à la probabilité que le code NOGA2 soit l sachant que les sections NOGA i, j , et k ont été sélectionnées au niveau précédent. Cette remarque s'applique également aux probabilités dans les niveaux inférieurs de l'algorithme.



Grâce à cette démarche, les modèles de prédictions sont moins complexes et prennent ainsi moins de temps à être entraînés. Ceci a permis de réduire le temps de calcul et d'améliorer la qualité de l'algorithme.

Validation croisée et hyperparamètres optimaux

La technique de validation croisée est un processus utilisé dans les projets d'apprentissage automatique afin de déterminer la meilleure combinaison d'hyperparamètres. On recherche donc la combinaison d'hyperparamètres qui permet de minimiser la « loss function ». Cette étape a été appliquée au modèle de prédiction des sections NOGA, les hyperparamètres optimaux sélectionnés ont ensuite été transférés sur les autres modèles. Ce choix a été effectué en raison de la capacité de calcul restreinte qui ne permet pas de faire une validation croisée pour l'ensemble des modèles.

Pour appliquer cette technique, l'ordinateur partitionne aléatoirement le « training set » initial en deux parties (un « validation set » et « training set » propre à la validation croisée-ci-après « training set_{val} »). L'image ci-dessous, représente la fonction dans R qui permet de créer une table de 24 combinaisons d'hyperparamètres. Chaque combinaison représente un modèle de prédiction des sections NOGA et sa performance associée.

```
hyper_table=expand.grid(eta=c(0.01,0.1),
                        max_depth=c(10,20),
                        gamma=c(1,2,0.5),
                        subsample=c(0.8,0.9),
                        min_child=c(5))
```

La fonction « logloss » permet d'optimiser le modèle en trouvant les meilleurs hyperparamètres. Elle mesure la distance entre la distribution empirique des classes réelles et la distribution des classes prédites.

$$F = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \cdot \ln(p_{ij})$$

Où N est le nombre d'éléments dans le « validation set », M le nombre de classes, y_{ij} une variable binaire (si l'élément i appartient à la classe j alors $y_{ij} = 1$, sinon $y_{ij} = 0$) et p_{ij} la probabilité (prédite par le modèle) que l'élément i appartienne à la classe j . On peut déduire de cette formule que la valeur de la « logloss » est plus faible lorsque les éléments sont classés correctement et avec une probabilité élevée.

Le processus de validation croisée suit la démarche suivante :

1. Pour chaque combinaison d'hyperparamètres, l'ordinateur entraîne un modèle à l'aide du « training set_{val} » de sorte que la valeur de la fonction de perte (training_logloss) soit minimale ;
2. Le modèle résultant est appliqué au « validation set » et la fonction de perte (validation_logloss) est mesurée ;
3. Les paramètres qui donnent la plus petite fonction de perte sur le « validation set » (validation_logloss) sont retenus pour entraîner le modèle final.

Le tableau suivant illustre les cinq meilleures combinaisons d'hyperparamètres obtenues par la validation croisée.

validation_logloss	training_logloss	Subsample	Gamma	Max_depth	Eta	Min_child
0.9989	0.2129	0.8	0.5	10	0.1	5
0.9990	0.1690	0.8	1	20	0.1	5
0.9990	0.2350	0.8	1	10	0.1	5
0.9999	0.2908	0.8	2	10	0.1	5
1.0003	0.1585	0.9	1	20	0.1	5

Les hyperparamètres optimaux retenus sont donc :

- Subsample: 0.8;
- Gamma: 0.5;
- Max_depth: 10;
- Eta: 0.1;
- Min_child: 5.

Ceux-ci sont directement intégrés dans la fonction d'entraînement « xgb.train » du package xgboost.

Exemple de résultats

Pour chaque niveau de l'algorithme, il est possible d'isoler les trois prédictions avec la plus haute probabilité. Ci-dessous se trouvent des exemples de résultats obtenus sur quelques descriptions d'activités avec les pourcentages associés (c'est-à-dire la probabilité que le modèle attribue à ce niveau NOGA pour la description d'activités insérée).

Prédictions au niveau des sections NOGA :

Text	Section NOGA	1st prediction	%	2nd prediction	%	3rd prediction	%
Réaliser tout type de matériel publicitaire, i	C	J	44.1	G	26.2	C	10.2
location et vente de machines et matériel d	N	N	49.8	C	34.4	G	12.1
toutes études en électromécanique et auto	C	C	91.1	G	4.99	N	0.92
La fabrication, la vente et la représentation	C	C	98.3	G	1.65	S	0.02
La société a pour but le commerce de produ	G	G	57.3	N	12.4	C	11.9
achat, vente et placement de distributeurs	C	G	65.7	C	27.6	N	3.24
importation et exportation de produits alim	G	G	98.5	C	0.8	A	0.33
conception et ingénierie de machines et un	C	C	59.2	G	34	M	2.12
exploitation d'une entreprise de dépannag	S	S	45.9	G	27.4	C	8.25
promouvoir l'art et la culture dans les doma	R	R	62.3	S	12.4	G	9.26

Prédictions au niveau des codes NOGA à 2 digits¹⁴:

Text	NOGA code 2D	1st prediction	%	2nd prediction	%	3rd prediction	%
Réaliser tout type de matériel publicitaire, i	18	47	59.8	46	10.6	18	4.87
location et vente de machines et matériel d	77	77	86.1	33	5.76	46	2.48
toutes études en électromécanique et auto	28	28	88.5	27	2.52	25	1.7
La fabrication, la vente et la représentation	25	25	96.1	26	1.01	32	0.49
La société a pour but le commerce de produ	46	47	56.1	10	15.2	46	7.8
achat, vente et placement de distributeurs à	10	47	29	28	24.6	46	20.6
importation et exportation de produits alim	46	46	87.61	47	12.2	45	0.18
conception et ingénierie de machines et un	28	46	67.4	28	5.24	20	3.94
exploitation d'une entreprise de dépannage	95	95	84.6	47	6.3	46	2.68
promouvoir l'art et la culture dans les doma	91	91	53.24	94	13.9	93	9.18

4.4. Evaluation et validation du modèle

Lorsque tous les modèles pour chaque phase ont été entraînés, la « predictive accuracy », la « precision », le « recall » et le « F1-score » ont pu être calculés grâce au test set comprenant 6'408 unités (pour rappel, nous traitons ici toujours des descriptions d'activités en français). La « predictive accuracy » donne une meilleure vue d'ensemble sur la performance globale de chaque modèle alors que les 3 autres mesures de qualité permettent de mettre en évidence les spécificités par catégorie.

Voici le tableau récapitulatif de la « predictive accuracy » lors de la prédiction des descriptions d'activités pour chaque catégorie. La « predictive accuracy » telle que présentée dans la première colonne donne la proportion d'éléments correctement prédits par rapport à tous les éléments du test set pour la première prédiction et correspond à la définition donnée précédemment. La deuxième colonne (« predictive accuracy 1st and 2nd prediction ») correspond à la proportion d'unités pour lesquelles les deux premières prédictions données par le modèle contiennent le bon code NOGA. De façon analogue, la troisième colonne (« predictive accuracy 1st, 2nd and 3rd prediction ») correspond à la proportion d'unités pour lesquelles les trois premières prédictions données par le modèle contiennent le bon code NOGA.

Predictive Accuracy	1st prediction	1st and 2nd predictions	1st, 2nd and 3rd predictions
Section NOGA	90%	96%	98%
2 digits	88%	93%	94%

¹⁴ Il est utile de noter à nouveau que les pourcentages affichés dépendent des choix effectués par l'algorithme au niveau précédent.

3 digits	63%	75%	81%
4 digits	50%	62%	68%
6 digits	48%	59%	64%

On peut déjà constater ici que le modèle est relativement performant jusqu'au niveau des 2 digits et que la qualité des résultats baisse à chaque passage à un niveau plus précis de la NOGA.

Afin de pouvoir détecter pour quelles catégories le modèle est moins performant, on peut se référer aux mesures de qualité « precision », « recall » et « F1-score ».

Pour rappel, la « precision » dans une catégorie mesure la proportion de True Positive (TP) sur le nombre de prédictions données par le modèle pour cette catégorie.

$$precision = \frac{\#TP}{\#TP + \#FP}$$

Concrètement, dans une catégorie de NOGA donnée, quel est le pourcentage d'unités pour lesquelles la prédiction de cette NOGA a été effectuée correctement. Une faible valeur de « precision » signifie que de nombreuses prédictions sont attribuées à cette NOGA alors qu'elles correspondent à un autre code NOGA.

Le « recall » dans une catégorie mesure le nombre de True Positive (TP) sur le nombre de vraies valeurs.

$$recall = \frac{\#TP}{\#TP + \#FN}$$

Concrètement, dans une catégorie de NOGA donnée, quel est le pourcentage d'unités pour lesquelles cette catégorie NOGA a été prédite par rapport au nombre réel d'observations dans cette catégorie. Une faible valeur de « recall » signifie que beaucoup de descriptions sont attribuées à tort à d'autres catégories NOGA et non à celle dont il est question.

Afin de faciliter la compréhension des valeurs du tableau ci-dessous, voici le détail des calculs pour la section NOGA C :

Nombre de valeurs réelles avec la section C dans le « test set » : 1119

1. Première prédiction:

TP: 1010 prédictions

FP: 148 prédictions

FN: 109 observations

$$\diamond Precision = \frac{1010}{(1010+148)} = \frac{1010}{1158} = 0.8721 \rightarrow 87.22\%$$

➔ Sur les 1158 prédictions du modèle, 87,22% ont été correctement prédites.

$$\diamond Recall = \frac{1010}{(1010+109)} = \frac{1010}{1119} = 0.9026 \rightarrow 90.26\%$$

➔ Sur les 1119 observations réelles, 90,26 % ont été correctement prédites par la première prédiction.

2. Première et deuxième prédiction:

TP: 1010 (nombre de premières prédictions) + 79 (nombre de deuxièmes prédictions égal à C, pour lesquelles la première prédiction n'est pas égale à C)

FP: 148 (nombre de premières prédictions) + 62 (nombre de deuxièmes prédictions qui sont également fausses)

FN: 30 observations

$$\diamond Precision = \frac{(1010+79)}{((1010+79)+(148+62))} = \frac{1089}{1299} = 0.8383 \rightarrow 83.83\%$$

➔ Sur les 1089 prédictions du modèle (première et seconde prédiction), 83,83% ont été correctement prédites.

$$\diamond Recall = \frac{(1010+79)}{((1010+79)+30)} = \frac{1089}{1119} = 0.9732 \rightarrow 97.32\%$$

➔ Sur les 1119 observations réelles, 97.32 % ont été correctement prédites par la première ou la deuxième prédiction.

3. Première, deuxième et troisième prédiction:

TP: 1010 (nombre de premières prédictions) + 79 (nombre de deuxièmes prédictions égal à C, dans lesquelles la première prédiction n'est pas égale à C) + 13 (nombre de troisièmes prédictions égal à C, dans lesquelles la première et la deuxième prédiction ne sont pas égales à C)

FP: 148 (nombre de premières prédictions) + 62 (nombre de deuxièmes prédictions qui sont aussi fausses) + 17 (nombre de troisièmes prédictions qui sont aussi fausses)

FN: 15 observations

$$\diamond Precision = \frac{(1010+79+13)}{((1010+79+13)+(148+62+17))} = \frac{1102}{1329} = 0.8295 \rightarrow 82.95\%$$

➔ Sur les 1102 prédictions du modèle (première, deuxième et troisième prédiction), 82,95% ont été correctement prédites.

$$\diamond Recall = \frac{(1010+79+13)}{((1010+79+13)+15)} = \frac{1102}{1119} = 0.9866 \rightarrow 98.66\%$$

➔ Sur les 1119 observations réelles, 98,66 % ont été correctement prédites par la première, la deuxième ou la troisième prédiction.

Le « F1-score » est la moyenne harmonique des deux mesures de qualité décrites ci-dessus. Voici le résultat au niveau des sections NOGA.

NOGA Section	Precision						Recall				F1-Score		
	Pred1 (%)	# Pred1	Pred1+ Pred2 (%)	# Pred 1+2	Pred1 + Pred2 + Pred3 (%)	# Pred 1+2+3	Rec1 (%)	Rec1 + Rec2 (%)	Rec1 + Rec2 + Rec3 (%)	# Obs.	F1-S1 (%)	F1-S2 (%)	F1-S3 (%)
A	98.11 ¹	371 ²	97.45 ³	392	96.39	416	84.65 ⁴	88.84 ⁵	93.26	430	90.88	92.95	94.8
B	95	20	95	20	95	20	100	100	100	19	97.44	97.44	97.44
C	87.22	1158	83.83	1299	82.95	1331	90.26	97.32	98.66	1119	88.71	90.07	90.13
D	98.25	57	98.36	61	98.36	61	90.32	96.77	96.77	62	94.12	97.56	97.56
E	100	83	94.62	93	92.71	96	90.22	95.65	96.74	92	94.86	95.13	94.68
F	90.62	320	86.71	346	85.2	358	92.36	95.54	97.13	314	91.48	90.91	90.77
G	89.94	1531	87.96	1636	86.64	1684	93.74	97.96	99.32	1469	91.8	92.69	92.55
H	92.61	257	89.82	275	88.15	287	91.54	95	97.31	260	92.07	92.34	92.5
I	92.96	142	90.26	154	87.12	163	88.59	93.29	95.3	149	90.72	91.75	91.03
J	89.75	322	85.34	348	83.75	363	94.14	96.74	99.02	307	91.89	90.68	90.75
K	96.36	220	93.22	236	90.73	248	93.39	96.92	99.12	227	94.85	95.03	94.74
L	88.17	93	84.85	99	83.33	102	91.11	93.33	94.44	90	89.62	88.89	88.54
M	91.98	349	86.8	394	84.35	409	89.66	95.53	96.37	358	90.81	90.96	89.96
N	88.58	499	84.5	555	81.97	588	90.2	95.71	98.37	490	89.38	89.76	89.42
O	63.64	33	64.1	39	64.29	42	75	89.29	96.43	28	68.85	74.63	77.15
P	89.9	198	85.65	230	83.9	236	88.12	97.52	98.02	202	89	91.2	90.41
Q	89.22	269	88.26	281	87.2	289	93.39	96.5	98.05	257	91.26	92.2	92.31

R	89.5	200	83.62	232	82.19	247	84.43	91.51	95.75	212	86.89	87.39	88.45
S	90.44	272	86.45	310	81.76	340	83.11	90.54	93.92	296	86.62	88.45	87.42
T	100	3	100	4	100	4	75	100	100	4	85.71	100	100
U	90.91	11	80	20	76.19	21	43.48	69.57	69.57	23	58.83	74.42	72.73

¹ Precision pour la section NOGA A : Dans les 371 cas où le modèle a prédit la section NOGA A (parmi les trois premières prédictions), 98% ont été prédits correctement en première prédiction.

² Correspond au nombre de fois où le modèle a prédit la section A parmi les trois premières prédictions.

³ Precision pour la section NOGA A : Dans les 392 cas où le modèle a prédit la section NOGA A (parmi les trois premières prédictions), 99% ont été prédits correctement dans une des deux premières prédictions.

⁴ Recall pour la section NOGA A : 85% des 430 observations du test set appartenant à la section NOGA A ont été prédites correctement en première prédiction.

⁵ Recall pour la section NOGA A : 89% des 430 observations du test set appartenant à la section NOGA A ont été prédites correctement parmi les deux premières prédictions.

Toutes les mesures de qualité pour les autres catégories (codes NOGA 2 digits, 3 digits, 4 digits et 6 digits) se trouvent dans un fichier excel nommé « FR_quality_metrics »¹⁵. Dès lors, il est plus facile d'identifier les catégories les plus complexes, telle que la section NOGA O qui n'a que 64% comme « precision », 75% comme « recall » et 69% comme « F1-score » lors de la première prédiction. Les classes pour lesquelles un nombre trop faible d'observation a été utilisé dans le « test set » peuvent aussi être considérée comme problématiques puisque seul un nombre restreint de prédictions a permis d'établir la valeur des mesures de qualité (la section NOGA T par exemple). La répartition des classes dans le « training set » est également disponible dans ce même document. Elle permet d'identifier les classes qui ne disposent pas d'un grand nombre d'observations lors de l'entraînement des modèles.

4.5. Modèle en allemand et italien

Pour les descriptions d'activités en allemand ou en italien, la même procédure de NLP est appliquée. Chaque jeu de données entraîne des modèles spécifiques pour chaque langue avec une taille globale de « training set » et de « test set » équivalente à celle de la langue française.

Pour la langue italienne nous obtenons des résultats concluants et les mesures de qualités ci-dessous nous confirment ces affirmations.

Predictive Accuracy	1st prediction	1st and 2nd predictions	1st, 2nd and 3rd predictions
Section NOGA	72%	85%	91%
2 digits	58%	70%	75%
3 digits	50%	62%	67%
4 digits	45%	56%	62%
6 digits	42%	53%	58%

Comme pour les descriptions d'activités en français, on peut se référer aux mesures de qualité « precision », « recall » et « F1-score », afin de pouvoir détecter pour quelles catégories le modèle est moins performant.

¹⁵ Q:\IOR\URD\06_02_NOGAuto\NOGAuto_APP\models\Model_evaluation

	Precision						Recall				F1-Score		
NOGA Section	Pred1 (%)	# Pred1	Pred1+ Pred2 (%)	# Pred 1+2	Pred1 + Pred2 + Pred3 (%)	# Pred 1+2+3	Rec1 (%)	Rec1 + Rec2 (%)	Rec1 + Rec2 + Rec3 (%)	# Obs.	F1-S1 (%)	F1-S2 (%)	F1-S3 (%)
A	97.03	1010	96.05	1038	94.83	1063	92.8	94.41	95.45	1056	94.87	95.22	95.14
B	40	5	42.86	7	50	12	16.67	25	50	12	23.53	31.58	50
C	58.44	308	54.31	615	52.08	795	36.73	68.16	84.49	490	45.11	60.45	64.44
D	50	8	54.55	11	37.5	16	33.33	50	50	12	40	52.18	42.86
E	71.43	7	42.86	14	47.06	17	22.73	27.27	36.36	22	34.49	33.33	41.02
F	71.64	804	67.79	975	65.94	1060	75.49	86.63	91.61	763	73.51	76.06	76.68
G	66.92	1977	64.33	2425	62.63	2617	78.1	92.09	96.75	1694	72.08	75.75	76.04
H	73.26	172	64.55	220	61.81	254	62.07	69.95	77.34	203	67.2	67.14	68.71
I	76.69	459	72.83	541	71.06	577	78.05	87.36	90.91	451	77.36	79.44	79.77
J	65.88	337	62.28	448	59.33	504	63.79	80.17	85.92	348	64.82	70.1	70.19
K	65.05	329	60.77	469	57.64	543	61.32	81.66	89.68	349	63.13	69.68	70.18
L	71.83	426	66.79	548	63.99	611	68.76	82.25	87.87	445	70.26	73.72	74.05
M	69.74	1821	64.32	2368	61.69	2574	77.06	92.42	96.36	1648	73.22	75.85	75.22
N	69.45	275	56.16	422	50.46	543	53.8	66.76	77.18	355	60.63	61	61.02
O	-	-	-	-	-	-	-	-	-	-	-	-	-
P	67.37	95	57.14	147	50.54	186	44.44	58.33	65.28	144	53.55	57.73	56.97
Q	80.63	568	73.46	682	70.05	728	85.45	93.47	95.15	536	82.97	82.27	80.69
R	47.66	107	46.67	180	45.7	256	30	49.41	68.82	170	36.82	48	54.93
S	68.53	483	61.13	620	56.53	727	69.25	79.29	85.98	478	68.89	69.04	68.21
T	-	-	-	-	-	-	-	-	-	-	-	-	-
U	65.22	23	53.57	28	53.57	28	42.86	42.86	42.86	35	51.73	47.62	47.62

Pour les descriptions d'activités en allemand, les mesures de qualités ne sont pas autant concluantes.

Afin que les descriptions d'activités en allemand puissent également être classifiées par le système NOGAuto, il a été décidé d'utiliser le logiciel de traduction DeepL pour que les descriptions soient traduites en français. Une description d'activités que l'algorithme de détection de la langue (décrit au chapitre 3.1) identifie comme étant écrite en allemand est traduite directement de l'allemand au français via DeepL puis traitée avec les modèles entraînés pour le français. Même si les résultats ne sont pas complètement satisfaisants, les mesures de qualités sont meilleures qu'avec la méthode utilisée initialement. Les résultats ne sont pas suffisants pour garder ce procédé sur une longue durée. NOGAuto peut cependant être utilisé pour le moment pour des descriptions d'activités en allemand mais avec beaucoup de précautions de la part de l'utilisateur. Un éventuel message de mise en garde peut être ajouté dans lors de l'utilisation si la description d'activités est en allemand.

Le tableau ci-dessous donne les mesures de la « predictive accuracy » pour les descriptions en allemand insérée dans le modèle pour la langue française.

Predictive Accuracy	1st prediction	1st and 2nd predictions	1st, 2nd and 3rd predictions
Section NOGA	55%	74%	83%
2 digits	37%	52%	60%
3 digits	25%	36%	42%

4 digits	19%	26%	31%
6 digits	16%	22%	27%

Comme pour les modèles français et italien, les mesures de qualité correspondantes ont été calculées.

NOGA Section	Precision						Recall				F1-Score		
	Pred1 (%)	# Pred1	Pred1+ Pred2 (%)	# Pred 1+2	Pred1 + Pred2 + Pred3 (%)	# Pred 1+2+3	Rec1 (%)	Rec1 + Rec2 (%)	Rec1 + Rec2 + Rec3 (%)	# Obs.	F1-S1 (%)	F1-S2 (%)	F1-S3 (%)
A	77.27	66	67	100	72.81	217	22.47	29.52	69.6	227	34.82	40.98	71.17
B	66.67	3	78.57	14	65	20	4.35	23.91	28.26	46	8.17	36.66	39.39
C	66.42	1784	61.62	2689	60.67	2919	63.74	89.13	95.27	1859	65.05	72.86	74.13
D	47.73	44	46.77	62	41.33	75	48.84	67.44	72.09	43	48.28	55.23	52.54
E	61.7	47	60.94	64	53.16	79	43.28	58.21	62.69	67	50.87	59.54	57.53
F	57.87	534	54.22	699	52.35	810	64.24	78.79	88.15	481	60.89	64.24	65.69
G	51.4	1817	48.82	2364	47.14	2550	74.3	91.81	95.62	1257	60.76	63.74	63.15
H	61.94	155	52.59	232	40.89	313	54.55	69.32	72.73	176	58.01	59.81	52.35
I	76.39	144	73.02	189	65.87	252	47.41	59.48	71.55	232	58.51	65.56	68.59
J	50.34	298	47.58	414	43.43	502	55.56	72.96	80.74	270	52.82	57.6	56.48
K	54.72	265	47.22	360	42.89	408	69.71	81.73	84.13	208	61.31	59.86	56.82
L	48.21	56	51.19	84	50.47	107	30.34	48.31	60.67	89	37.24	49.71	55.1
M	57.06	354	54.46	628	50.64	859	35.56	60.21	76.58	568	43.81	57.19	60.97
N	29.17	504	25.32	873	24.22	1115	44.55	66.97	81.82	330	35.26	36.75	37.38
O	22.73	22	27.27	33	30.95	42	8.93	16.07	23.21	56	12.82	20.22	26.53
P	55.69	167	52.88	208	51.24	242	45.81	54.19	61.08	203	50.27	53.53	55.73
Q	77.14	175	69.76	248	68.86	273	47.7	61.13	66.43	283	58.95	65.16	67.62
R	33.04	224	34.98	303	31.18	372	43.02	61.63	67.44	172	37.38	44.63	42.64
S	53.63	179	45.79	297	40.41	438	37.5	53.12	69.14	256	44.14	49.18	51.01
T	0	1	0	1	0	2	0	0	0	15	NA	NA	NA
U	-	-	-	-	-	-	-	-	-	-	-	-	-

A l'avenir plusieurs pistes peuvent être explorées pour tenter de résoudre ce problème. Une meilleure infrastructure peut permettre d'entraîner des modèles avec un « training set » beaucoup plus conséquent et donner de possibles meilleurs résultats. Un traitement spécifique à la langue allemande pour le NLP est également une éventualité à envisager. Une dernière piste est un entraînement de modèles allemands avec des descriptions d'activités traduites de l'allemand au français pour éviter des biais de vocabulaire (une traduction via DeepL ne fournit pas nécessairement un vocabulaire similaire à des descriptions d'activités formulées directement par une personne francophone).

5. Révision

5.1. Processus de révision

Le processus de révision doit encore être élaboré d'entente avec la section METH. Il existe cependant quelques points qui peuvent être relevés sur les possibilités actuelles de révision.

1. Possibilité de réentraîner un modèle avec des données plus actuelles car aucune prédiction de NOGAuto n'est insérée telle quelle dans le registre. On évite ainsi que l'algorithme s'alimente lui-même. La façon de procéder à un réentraînement

correspondrait en tous point à celle décrite dans ce document. Les mesures de qualité seraient adaptées aux nouveaux modèles.

2. Dès le moment où l'infrastructure le permettrait, les modèles pourraient être entraînés avec une plus grande quantité de données. L'utilisation de l'intégralité des données à disposition pour procéder à l'entraînement ainsi qu'au test devrait nous permettre d'augmenter la qualité des résultats. Par ailleurs une méthode d'oversampling pourrait être appliquée pour que les catégories NOGA sous-représentées dans le jeu de données initial soient mieux considérées lors de l'entraînement. Les mesures de qualité seraient alors recalculées avec un nouveau test set dont la répartition respectera celle utilisée actuellement (i.e. 20% du jeu de données initial).

6. Déploiement

6.1. Déploiement partiel

Afin d'utiliser au mieux le produit qui a été développé sans introduire de biais dans la classification et dans l'optique d'attendre une meilleure infrastructure, il a été décidé de procéder à une mise en production partielle de NOGAuto. Cette décision est appuyée également sur les besoins actuels en matière de codification NOGA.

Cette mise en production partielle est caractérisée par la façon d'utiliser l'algorithme de NOGAuto. Il est donc envisagé d'utiliser le système à des fins de contrôle de qualité de listes. Dans le REE et dans d'autres registres, il arrive que la NOGA soit mal codée. Un contrôle systématique des codes NOGA est difficile à envisager en raison du grand nombre de données à traiter. En insérant des listes dans NOGAuto (description d'activités et code NOGA existant), il est possible d'effectuer un pré-contrôle. Le système est en mesure de nous dire si :

1. Il est d'accord avec le code NOGA existant ;
2. Il est d'accord avec un certain nombre de digits du code NOGA existant ou non.

Ainsi il est possible d'établir une priorité dans le traitement des listes de contrôle afin d'améliorer la qualité. (par exemple en traitant dans un premier temps tous les éléments de la liste pour lesquels NOGAuto n'est pas d'accord avec la section NOGA existante, puis tous les cas pour lesquels NOGAuto n'est pas d'accord avec le code NOGA 2 digits existant, etc).

Pour éviter d'introduire un biais dans l'esprit du collaborateur qui doit traiter ces listes, on peut décider que NOGAuto ne retourne que deux valeurs : « d'accord » ou « pas d'accord », ainsi le collaborateur n'est pas influencé par le code NOGA que NOGAuto aurait lui-même attribué à cette description d'activités. Il est également possible de proposer au collaborateur les trois codes NOGA les plus probables selon le système.

Cette façon séquentielle de procéder est également appuyée sur le fait que l'algorithme est mieux entraîné pour les parties hautes de l'arborescence NOGA, il y a donc plus de chances que la prédiction soit correcte pour une section NOGA que pour un code NOGA à 6 digits.

Cette fonctionnalité a déjà été testée avec succès. Actuellement, la machine prend beaucoup de temps pour traiter une liste, mais les résultats sont concluants et simplifient grandement le traitement de listes pour les collaborateurs concernés.

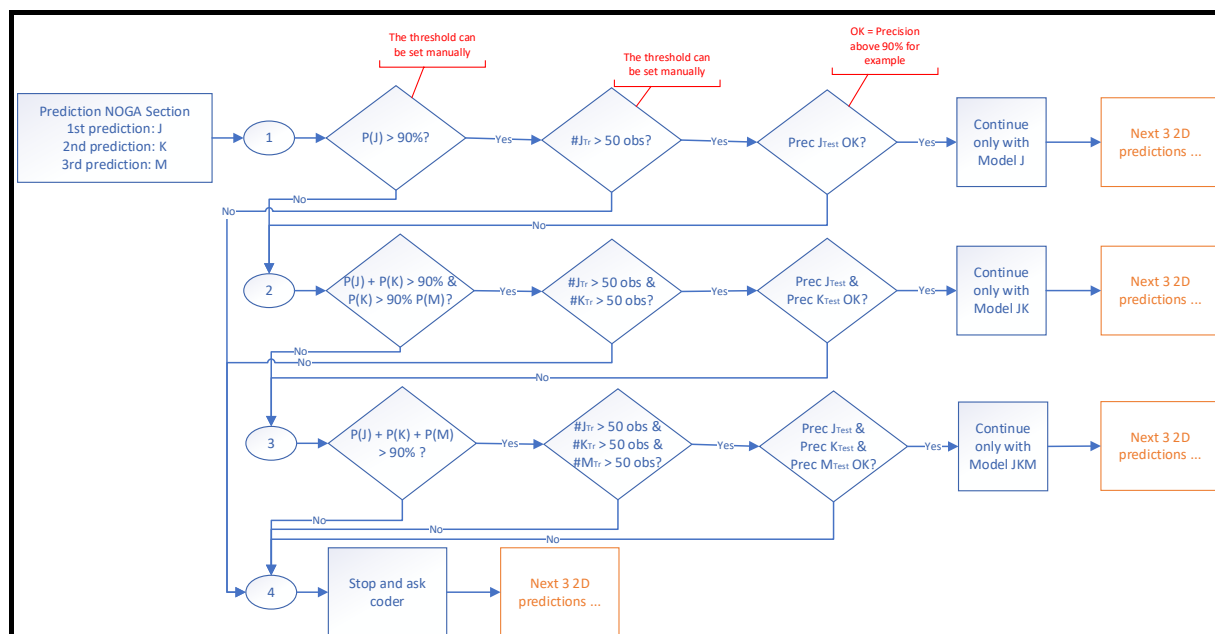
En perspective de la révision de la NOGA prévue pour 2025, il est également intéressant de pouvoir utiliser la fonctionnalité de vérification de NOGAuto pour les travaux préparatoires.

6.2. Déploiement futur

L'application doit permettre à l'utilisateur de pouvoir interagir avec les modèles sans devoir écrire ou exécuter une ligne de code dans R. Le fonctionnement de l'algorithme diffère de celui pour le contrôle de listes. Pour éviter de laisser l'algorithme tourner en roue libre dans des classes pour lesquelles la qualité est moindre, un système de seuils peut être mis en place. Ce système permet au codeur d'interagir directement avec NOGAuto lorsque nécessaire. La

valeur exacte des seuils reste encore à déterminer avec la section METH et fait également partie des raisons pour lesquelles le choix d'une mise en production partielle a été effectué.

Si l'algorithme remplit les critères des différents seuils, il passe de lui-même à l'étape suivante, sinon, il demande l'avis de l'utilisateur afin de pouvoir continuer les calculs. Le diagramme ci-dessous illustre ce système de fonctionnement de l'algorithme à l'étape de prédiction NOGA 2 digits :



Afin que les collaborateurs puissent utiliser le produit d'une façon plus intuitive, un « User Interface (UI) » a été développé. Grâce au package « Shiny », spécialement conçu pour le développement d'applications pour la visualisation de projet d'apprentissage automatique, un prototype a été créé. Celui-ci pourrait ressembler à la future application. Néanmoins, afin que le produit soit viable sur le long terme, il faut trouver une interface plus adéquate de déploiement au sein de l'OFS.

L'image ci-dessous illustre le prototype du UI.

Prédiction du code NOGA

Choisir la langue
fr

Insérer les variables nécessaires

Saisissez la description d'activité de l'entreprise

Langue détectée:
French

Si la langue est incorrecte, veuillez choisir:
☐ Français ☐ Allemand ☐ Italien

Search

Première prédiction

Train model

Final code

Deuxième prédiction

Train model

Final code

Troisième prédiction

Train model

Final code

Traduction de texte

Nettoyage de texte

Feedback:

Modification du code par le codeur:
Choix de la catégorie NOGA

Section NOGA 2 digits 3 digits
4 digits 5 digits

Modify code

Ecrivez un commentaire ci-dessous

Envoyer commentaire

Close window

Afin de contrôler les accès à l'application, une phase d'authentification a été mise en place, où chaque collaborateur entre son identifiant afin d'utiliser NOGAuto. Ceci permet de contrôler les accès à l'application mais sert également de sécurité. A l'avenir, il est prévu que l'application offre différentes options selon le degré d'expertise de l'utilisateur.

Voici ci-dessous une illustration du fonctionnement de l'algorithme avec seuils. Les seuils ont été fixés arbitrairement uniquement pour illustrer cette façon de procéder. La description d'activités, « Exploitation d'une entreprise de menuiserie et charpente » correspondant au code NOGA à 4 digits 1623 est insérée dans le système. Lorsque le collaborateur introduit la description d'activités dans l'espace dédié, la langue est automatiquement détectée.

Insérer les variables nécessaires

Saisissez la description d'activité de l'entreprise

Exploitation d'une entreprise de menuiserie et charpente

Langue détectée:
French

Si la langue est incorrecte, veuillez choisir:
☐ Français ☐ Allemand ☐ Italien

Search

La langue détectée automatiquement par l'algorithme est le français. Si le collaborateur est d'accord avec le système, il clique sur le bouton "Search", ce qui déclenche le processus de prédiction. Si le collaborateur n'est pas satisfait de la langue détectée, il a la possibilité d'en choisir une autre.

Dans le cas d'une description d'activités identifiée comment étant rédigée en allemand, le système traduit automatiquement via DeepL la description de l'allemand au français.

Lorsque le programme a effectué les différentes tâches, les premières prédictions sont présentées comme ci-dessous.

Première prédiction

1623
Fabrication de charpentes et d'autres menuiseries

54.6 %

Train model

Final code

Deuxième prédiction

1622
Fabrication de parquets assemblés

42.7 %

Train model

Final code

Troisième prédiction

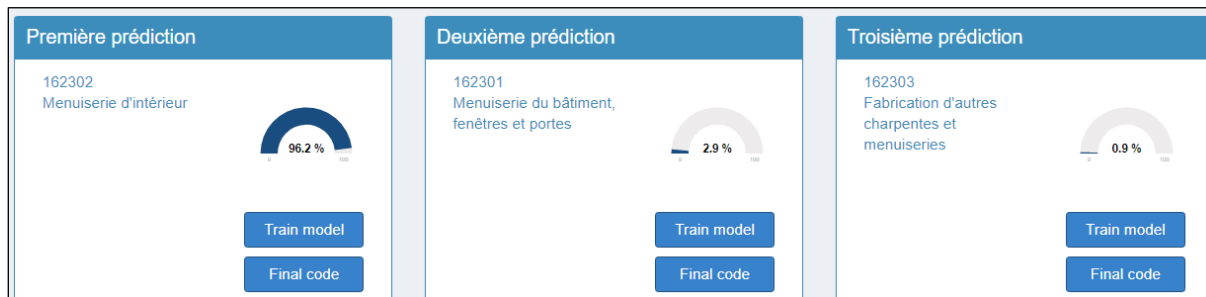
1629
Fabrication d'objets divers en bois; fabrication d'objets en liège, vannerie et sparterie

1.1 %

Train model

Final code

Dans cet exemple, grâce aux paramétrages de tous les seuils, NOGAuto a réussi à aller de lui-même jusqu'aux 4 digits. A ce stade, l'utilisateur lui indique quel est, selon lui, le bon code NOGA à 4 digits. En cliquant sur « train model », il permet à NOGAuto de continuer le processus de prédiction vers les codes finaux à 6 digits.



L'utilisateur peut sélectionner le code qui lui semble être correct en cliquant sur « final code ». Ce code est enregistré dans une base de données « prédictions choisies par le codeur » pour un monitoring des résultats par l'équipe NOGA. Si l'utilisateur n'est pas convaincu par les résultats donnés par NOGAuto, il peut changer le code dans la section feedback. Il y trouve tous les codes de la nomenclature des entreprises répartis par catégories.

Modification du code par le codeur:
Choix de la catégorie NOGA

Section NOGA 2 digits 3 digits 4 digits 6 digits

Choix du code NOGA Choisir... ▾

Modify code

Si l'utilisateur actionne le bouton « Modify code », une commande d'action est envoyée au système et NOGAuto reprend le code que l'utilisateur lui a donné pour faire de nouvelles prédictions dans la catégorie indiquée. Il est prévu de référencer les cas où l'utilisateur n'as pas choisi un des codes suggérés par NOGAuto.

Voici un exemple :

Traduction du texte	
Texte original	Die Gesellschaft bezweckt den Betrieb einer Yoga-, Tanz- und Bewegungsschule, die Organisation von Yogaferien, Ausbildungen und Workshops sowie die Forschung im Gesundheitsbereich.
Texte traduit	La société a pour but l'exploitation d'une école de yoga, de danse et de mouvement, l'organisation de vacances de yoga, de formations et d'ateliers ainsi que la recherche dans le domaine de la santé.

L'utilisateur peut également voir de quelle façon le texte a été nettoyé et quels mots ont été sélectionnés pour la partie NLP.

Nettoyage du texte	
Texte original	La société a pour but l'exploitation d'une école de yoga, de danse et de mouvement, l'organisation de vacances de yoga, de formations et d'ateliers ainsi que la recherche dans le domaine de la santé.
Texte nettoye	ecol yog dans mouv organis vacanc format ateli recherch domain sant

7. Prochaines étapes

Les prochaines étapes du projet doivent être définies avec la section METH. Voici une liste non-exhaustive des prochaines pistes qui pourraient être explorées :

- Réentraînement des modèles avec plus de données (idéalement sur toutes les données disponibles) sur une infrastructure adaptée ;
- Compte tenu du fait que le jeu de données réduit n'est pas un simple échantillon aléatoire : analyse de l'influence de la réduction, par exemple sur les probabilités de prédiction et les mesures de qualité ;
- Mise en place d'un processus ainsi que d'une planification pour la révision des modèles ;
- Améliorations du modèle en allemand ;
- Utilisation d'autres variables pour la prédiction des codes NOGA (forme juridique, nombre d'employés, etc) ;
- Détermination des seuils pour l'utilisation de l'application de prédiction ;
- Introduction d'une variable indiquant si le code NOGA a été déterminé à l'aide de NOGAuto ou non ;
- Gestion du monitoring et des retours des collaborateurs ;
- Déploiement à plus grande échelle de l'application ;
- Utilisation de la même méthodologie dans d'autres contextes (par exemple la nomenclature des professions).

8. Révision du projet

Comme il a été opté de passer en production partielle pour le système NOGAuto, les critères définis par la section METH pour un passage en production sont partiellement remplis.

	Critères	Lien
1.	Allgemeiner nachvollziehbarer Beschrieb des Pilotprojekts.	La création d'un programme qui prédit automatiquement les codes NOGA des entreprises utilisant des techniques de machine learning (ML) avec une qualité de codage qui est égale ou dépasse celle du codage manuel utilisé actuellement. La qualité de codage sera évaluée selon les critères définis par le service des nomenclatures.
2.	Ziele und Rahmenbedingungen des Pilotprojekts, sowie eine Beurteilung der Zielerfüllung sowie Beitrag zur Dateninnovationsstrategie.	Création / Mise à disposition d'un programme qui, grâce aux techniques de machine learning, puisse associer de façon automatique les codes NOGA aux entreprises enregistrées dans le REE (BUR) avec une qualité de codage égale ou meilleure à la codification manuelle effectuée actuellement. Les bénéfices sont: • Vérification automatique des codes NOGA des entreprises du REE ; • Standardisation de la codification ; • Réduction de l'importance de l'interprétation ; • Réduction du temps consacré à la codification; • Amélioration de la qualité du codage.
3.	Welches ist der Mehrwert in Bezug auf die Problemstellung?	Actuellement, l'attribution et le contrôle des codes NOGA se font manuellement, ce qui amène à des divergences dans l'interprétation des activités économiques des entreprises, ce qui crée une certaine instabilité dans la qualité de la codification des codes NOGA. Ceci pourrait engendrer une baisse de qualité dans les différents résultats statistiques. La qualité de la codification NOGA des entreprises enregistrées dans le registre des entreprises et des établissements (REE ou BUR en allemand) a un impact direct sur les résultats des statistiques structurelles, conjoncturelles et de synthèse qui concernent les entreprises. Cet outil d'aide à la codification sera amélioré au fil du temps grâce aux inputs des collaborateurs.

4.	Welche ist die globale Relevanz?	La Nomenclature générale des activités économiques (NOGA) est un élément essentiel de la production statistique à l'OFS. La qualité de la codification NOGA des entreprises enregistrées dans le registre des entreprises et des établissements (REE ou BUR en allemand) a un impact direct sur les résultats des statistiques structurelles, conjoncturelles et de synthèse qui concernent les entreprises. Celles-ci dépendent d'une codification NOGA stable, contrôlée et de qualité des unités du REE. Dans un souci de décharge des entreprises et d'amélioration continue de la codification des entreprises, le projet vise à automatiser l'attribution des codes d'activité économique aux entreprises, en se basant, dans un premier temps, sur les informations déjà à disposition au sein de l'OFS. Il est envisagé d'utiliser la méthodologie sur d'autres nomenclatures telle que celle des professions. Un premier contact a déjà été fait avec les sections AES, LOHN et POP. Les premiers résultats sont encourageants.
5.	Wie wird die Referenzgrösse („Baseline“), der Ist-Zustand (vor dem Pilotprojekt), gemessen, mit welchem die Resultate des Pilotprojekts verglichen werden („Baseline“ Vergleich)?	Aujourd'hui, le collaborateur est seul face à son interprétation, mais grâce au projet pilote une harmonisation du codage pourrait être assurée grâce à ce système de support. De plus le collaborateur expérimenté pourrait aider NOGAuto à s'améliorer. A déterminer avec la section METH.
6.	„Lessons learned“ im Verlauf der Pilotprojektphase.	Tout au long du projet, nous avons pu revoir les techniques qui ont été utilisées et qui ont pu être améliorées.
7.	Selbständiges Arbeiten des Pilotprojektteams im Rahmen des «Coachings» («training on the job»)?	Le coaching a été assuré par le professeur Diego Kuonen ainsi que la section METH.
8.	Automatisierung (Effizienzgewinn)?	Réduction du choix de 794 codes NOGA à une prédiction de 3 possibilités. Ceci réduit fortement le temps de recherche pour un collaborateur.
9.	Detaillierter Beschrieb der verwendeten Daten, Datenquellen und deren Qualität (z.B. Abdeckung, Messwertfehler, Bearbeitungsfehler), sowie deren Verfügbarkeit und Vergleichbarkeit/Stabilität (mittelfristig und langfristig) und Verknüpfbarkeit. Auswirkungen auf die Metadaten?	Tous les détails par rapport aux données utilisées peuvent être trouvés dans le chapitre « Compréhension des données »
10.	Detaillierter und reproduzierbarer Beschrieb des «Feature Engineering» (Variablen-konstruktion).	Chapitre «Feature Engineering»

11.	Detaillierter Beschrieb der angewandten Methoden und deren Stärken und Schwächen (z.B. Stabilität/Robustheit der Methoden, Effizienz der Methoden bezüglich Prädiktion/Prediction und Erklärung/Schätzung, Vorgehen zur Auswahl der allfälligen „Hyperparameter“). Gibt es Alternativen oder Verbesserungspotenzial?	Chapitre «Modélisation»
12.	Detaillierter und reproduzierbarer Beschrieb der Methoden zur Validierung der Resultate inklusive Gütemass(e) in Bezug auf die Prädiktion/Prediction (z.B. Verallgemeinerungsfehler, „generalisation (prediction) error“) und die Erklärung/Schätzung (z.B. Modellfehler und Modellpräzision, „goodness of fit“), Vorgehen bei deren Berechnung(en). Gibt es Alternativen oder Verbesserungspotenzial?	Chapitre « Modélisation »
13.	Vorgehensprozess für die Qualitätssicherung, das Monitoring der angewandten Methoden (inklusive Häufigkeit/Frequenz der Aktualisierung der Methoden, „recalibration“ und „model change management“) und das Monitoring der verwendeten Daten, wenn diese Methoden in der statistischen Produktion angewandt werden. Beschrieb der relevanten Qualitätsmerkmale des Produktionsprozesses, z.B. sind die Resultate vergleichbar über die Zeit?	Grâce aux inputs des utilisateurs de l'application, l'équipe en charge de la maintenance du système pourra réentraîner les modèles et pourra étudier les feedbacks des experts. Ces derniers permettront à l'équipe de maintenir le système à jour, afin que ce dernier ne devienne obsolète, car il ne correspond plus aux attentes des utilisateurs. Il y'aura un réentraînement régulier des modèles. Ce processus sera défini avec la section METH et sera détaillé ainsi que finalisé dans un prochain rapport.
14.	Detaillierter Beschrieb der produzierten Daten/Resultate und deren Qualität (z.B. Abdeckung, Messwertfehler, Bearbeitungsfehler), sowie deren Verfügbarkeit und Vergleichbarkeit/Stabilität (mittelfristig und langfristig) und Verknüpfbarkeit. Auswirkungen auf die Metadaten?	Donner des prédictions des codes NOGA et une confirmation de la langue utilisée par l'entreprise pour la description d'activités. La mesure de qualité principale (avec celles détaillées dans ce rapport), est la satisfaction du collaborateur qui utilise le système. Si ce dernier estime que le système devient contraignant, une analyse approfondie et un ajustement des modèles devrait être envisagé. Il n'y a aucune influence sur la nomenclature. A déterminer avec la section METH.
15.	Automatisierung (Effizienzgewinn) in Bezug auf die Entwicklung des Pilotprojekts?	En proposant des prédictions aux collaborateurs nous réduisons le biais d'interprétation de ces derniers. Ceci pourrait avoir un impact significatif sur la qualité des classifications des entreprises dans la NOGA. Le gain se fait au niveau de la réduction du temps de codification et dans le contrôle de listes. A déterminer avec la section METH.

16.	Allgemeine Begründung des Transfers in die Produktion.	Ceci nous permettrait de déployer le système et permettre aux différents collaborateurs de l'utiliser et de l'évaluer. Grâce à leur évaluation, le produit pourra être amélioré avec le temps.
17.	Zeitpunkt des Transfers. Ist dieser realistisch, auch in Bezug auf die nötigen Ressourcen, und mit Revisionsvorhaben (falls vorhanden) abgestimmt?	Le projet est déjà mature pour le traitement de listes. Une application a été prédéveloppée et pourrait être opérationnelle lorsque les niveaux des seuils auront été déterminés.
18.	Sind juristische Anpassungen nötig? Falls ja, Stellungnahme von RD.	Non.
19.	Kann der Datenschutz gewährleistet werden? Stellungnahmen DSK und ISBO.	Cela dépend de la solution apportée. Les données de description d'activités économiques ne sont pas des données sensibles.
20.	Sind Anpassungen des Produktionsprozesses nötig? Konzeptueller Beschrieb des Transferprozesses.	Connaître comment nous pouvons déployer le projet au sein de l'office. En cours de planification. A déterminer avec la section METH.
21.	Ist die Methode skalierbar (auf grössere Datenmengen, respektive unter unterschiedlichen Rahmenbedingungen anwendbar)?	Oui, ce projet devrait être applicable à d'autres types de nomenclatures, comme celle des professions. Cependant, l'infrastructure allouée pour un plus grand nombre de données doit être assurée. Ce qui n'est pas encore le cas. A déterminer avec la section METH.
22.	Sind die angewandten Methoden wiederverwendbar?	Oui, les méthodes peuvent être réutilisées sur le même type de données (descriptions d'activités).
23.	Einbettung der entsprechenden Methoden, Resultate und Daten in der Produktionsumgebung?	Application web comme le KUBB (modernisation de l'outil de codification KUBB)
24.	Automatisierung (Effizienzgewinn) in Bezug auf die Produktion?	En proposant des prédictions aux collaborateurs nous espérons réduire le biais d'interprétation des collaborateurs. Ceci aura un impact significatif sur la qualité des classifications des entreprises dans la NOGA. De plus, un gain de temps dans le traitement des listes pourra être apporté.
25.	Sind Anpassungen der bestehenden Produktions-Metadaten nötig? Falls ja, wie wird dies umgesetzt und kommuniziert?	Les métadonnées seront adaptées en fonction des révisions de nomenclatures, c'est un procédé qui engendre des publications officielles. A déterminer avec la section METH.
26.	Ist eine Revision bereits publizierter Resultate nötig?	Aucune publication prévue. Utilisation à l'interne.

27.	Wie wird mit Zeitreihen umgegangen? Braucht es Anpassungen, respektive eine Retropolation?	Il y aura un réentraînement régulier (périodicité à déterminer) des modèles. En effet, les collaborateurs pourront sélectionner les codes que le système NOGAuto leur donnera et ces derniers vont être introduits dans le jeu de données après vérification de l'équipe responsable du monitoring. Grâce aux jeux de données actualisés, les nouveaux modèles entraînés devraient devenir toujours plus précis. Bien sûr, le monitoring de l'équipe en charge de la maintenance assurera la qualité des données ajoutées au training set. A déterminer avec la section METH
28.	Beschrieb der verwendeten Software/Pakete/Bibliotheken (inklusive Versionsnummern) und Abklärung, ob diese gegebenenfalls in der Produktion eingesetzt werden können (falls nicht über Schnittstellen möglich).	Les packages sont consultables dans l'annexe N.8.
29.	Auswirkungen auf die Infrastruktur/Softwarelandschaft? Falls ja, konzeptueller Beschrieb (Skizze) der Infrastrukturanpassungen (gilt auch für Schnittstellen).	A déterminer avec la section METH
30.	Sind Anpassungen des Publikationsprozesses nötig? Falls ja, Stellungnahme von DIAM.	Non, il n'y a pas de communication. Outil utilisé pour les besoins internes.
31.	Kommunikation: wie ist diese im BFS und gegenüber Nutzern (u.a. FEDESTAT, REGIOSTAT, KORSTAT) sichergestellt und auf potenzielle Änderungen ausgerichtet?	Dans une première étape, non, mais quand le produit sera de plus en plus performant, nous examinerons les différentes possibilités d'extension.
32.	Auswirkungen auf die nötigen Personalressourcen?	Performance dans le codage et outil peut s'avérer utile lors des cycles de révision de la NOGA. Temps de codage réduit au profit d'autres activités.
33.	Besteht die Notwendigkeit einer zusätzlichen Ausbildung des Produktionsteams?	Non, l'outil est destiné aux personnes ayant les compétences adéquates (codeurs NOGA)
34.	Finanzielle Auswirkungen?	Idem au système KUBB, internalisation des compétences de développement → réduction coûts.

9. Annexes

Annexe N.1: Evaluation du prototype

Fiche d'évaluation du prototype (phase d'introduction)	
Nom du logiciel	NOGAuto
Date d'évaluation	
Nom évaluateur	
Environnement test	
Points généraux	
Nombre de cas traités	
Langue utilisée	☐ Allemand ☐ Français ☐ Italien ☐ Anglais
Qualité de la prédiction	
Est-ce que la prédiction est satisfaisante ?	1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐
Indiquez la raison de votre insatisfaction concernant la qualité de la prédiction <i>(uniquement si 1 ou 2) (facultatif)</i>	
Plus-value de la prédiction	
Est-ce que la prédiction fournie est une plus-value ?	1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐
Expliquez votre note <i>(facultatif)</i>	
Layout	
Masque de saisie	1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐
Performance	
Rapidité de l'affichage des résultats	1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐
Estimez-vous qu'il y a des améliorations à apporter au prototype ? <i>(Veuillez indiquer le(-s) point(-s) concerné(-s))</i>	

Annexe N.2: Vorgehen für einen möglichen Transfer in die Produktion



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Dateninnovationspilotprojekte

Vorgehen für einen möglichen Transfer in die Produktion

Mandant GL
Autoren Daniel Kilchmann (DKi), Jean-Pierre Renfer (JPR) R/METH,
Diego Kuonen (DKu) Statoo Consulting
Klassifikation Intern
Stand Unterbreitung GL
Datum 15.01.2020

Versionen

Datum	Version	Änderung	Autor
21.06.2019	0.1	Erste Version, basierend auf Diskussion, JPR, DKu, DKi vom 05.06.2019	D. Kilchmann
01.07.2019	0.2	Überarbeitete Version nach Diskussion DKi, JPR vom 28.06.2019	D. Kilchmann
17.07.2019	0.3	Überarbeitete Version nach Diskussion DKi, DKu vom 15.07.2019	D. Kilchmann
24.07.2019	0.4	Überarbeitete Version nach Diskussion DKi, DKu vom 22.07.2019	D. Kilchmann, D. Kuonen
24.09.2019	0.5	Überarbeitete Version nach Diskussion DKi, DKu vom 24.09.2019	D. Kilchmann, D. Kuonen
07.10.2019	0.6	Überarbeitete Version nach Diskussion JPR, DKi vom 04.10.2019: Kriterien werden auf einige wenige reduziert, die Restlichen werden in den Anhang verschoben und als Beurteilungselemente bezeichnet	D. Kilchmann
16.10.2019	0.7	Überarbeitete Version nach Diskussion JPR, DKi, DKu vom 15.10.2019	D. Kilchmann
18.10.2019	0.8	Formattierung und leichte Anpassung durch DKu	D. Kilchmann, D. Kuonen
25.11.2019	0.9	Einarbeitung der Kommentare der AG NDS. «Stand der Pilotprojekte» wird in separaten Anhang verschoben. Hinzufügen des Geltungsbereichs. Diskussion JPR, DKi, DKu	D. Kilchmann, D. Kuonen, J.-P. Renfer
28.11.2019	0.91	Vorschlag für letzten(?) offenen Punkt bezüglich externer Pilotprojektanträge (Punkt 3. in 2.2). Hinzufügen einer Evaluation der Beurteilung der Pilotprojektanträge in die Empfehlung zum Transfervorhaben	D. Kilchmann

Datum	Version	Änderung	Autor
16.12.2019	0.92	Input DKu leicht angepasst + Inputs JPR	D. Kilchmann
19.12.2019	0.93	Diskussion DKu und DKi	D. Kilchmann
20.12.2019	0.94	Diskussion JPR und DKi	D. Kilchmann
27.12.2019	0.95	Inputs DKu akzeptiert, Kommentare gelöscht	D. Kilchmann
14.01.2020	0.96	Letzte Formatierung, Beurteilungsstufen Text hinzugefügt.	D. Kilchmann
15.01.2020	0.97	Letzte Schreibfehler korrigieren.	D. Kilchmann, D. Kuonen

Inhaltsverzeichnis

1	Einführung	3
1.1	Rahmenbedingungen und Ziel dieses Dokuments.....	3
2	Ausgangslage	4
2.1	Dateninnovationsstrategie 1.0	4
2.2	Auswahlverfahren für die Pilotprojekte der Dateninnovationsstrategie 1.0	4
2.3	Kommunikation mit AG DS und GL	5
2.4	Geltungsbereich	5
3	Vorgehen für einen möglichen Transfer in die Produktion	6
3.1	Beurteilungsprozess	6
3.2	Beurteilung	6
4	Anhang: Zusätzliche Beurteilungselemente	8

1 Einführung

In Bezug auf die herkömmliche statistische Produktion bestehen allgemein anerkannte Vorgehensweisen (u.a. GAMSO, GSBPM) und Qualitätskriterien des BFS («Manuel d'assurance qualité pour la production de statistiques à l'OFS»), die darauf abzielen, qualitativ hochstehende Grundlagen für die politische Entscheidungsfindung in der Schweiz zu liefern. Im Bereich der Dateninnovation müssen solche Vorgehensweisen festgelegt, respektive erweitert werden, da sie durch die erwähnten Hilfsmittel entweder nur teilweise oder nicht detailliert genug beschrieben sind.

1.1 Rahmenbedingungen und Ziel dieses Dokuments

Dieses Dokument bezieht sich auf den Wissenstand vom 28.06.2019 und den eingegangenen Pilotprojektberichten. Somit wurde es mit dem Hintergrund der Pilotprojekte der Dateninnovationsstrategie 1.0 erarbeitet und das Vorgehen basiert unter anderem auf den Auswahlkriterien für die Pilotprojekte und den Erfahrungen aus den Pilotprojekten.

Aufgrund potenzieller Änderungen der Rahmenbedingungen sollte dieses Dokument periodisch überprüft und nötigenfalls angepasst werden.

Es gilt folgender Grundsatz:

Pilotprojekte der Dateninnovationsstrategie sind experimentelle Statistiken (inklusive einzelner Statistikproduktionsschritte) und werden mit einem leicht erkennbaren Logo markiert, solange sie von der GL nicht als Standardprodukte für die entsprechende Statistikproduktion freigegeben wurden.

Experimentelle Statistiken umfassen somit nicht nur Output Produkte wie eine Statistik, sondern auch Aktivitäten, die Teile einzelner Produktionsschritte sind, z.B. Validierung, Codierung, Klassifizierung etc. Beispielsweise mittels komplementärer Analysemethoden aus dem „Machine Learning“ erstellte Daten, welche wiederum als Input für die Verbesserung und/oder Erweiterung einer bestehenden Statistik genutzt werden können.

2 Ausgangslage

2.1 Dateninnovationsstrategie 1.0

Die Dateninnovationsstrategie 1.0 hatte zum Ziel, Erfahrungen mit dem Umgang von komplementären Methoden zu sammeln, das Potenzial dieser Methoden zu eruieren, sowie die Infrastruktur, die Fähigkeiten der Mitarbeiter und die Organisation zu testen. Der Übergang in die Produktion wurde bei der Auswahl der Pilotprojekte zwar berücksichtigt, aber es war von vornherein klar, dass nicht alle Pilotprojekte (oder Teile davon) zum gleichen Zeitpunkt in die ordentliche statistische Produktion überführt werden können.

2.2 Auswahlverfahren für die Pilotprojekte der Dateninnovationsstrategie 1.0

Die Pilotprojekte wurden durch Jean-Pierre Renfer, Daniel Kilchmann und Diego Kuonen beurteilt. Die Vorschläge wurden Bertrand Loison unterbreitet und am 21.02.2018 von der GL gutgeheissen.

Die Beurteilung wurde durch eine unabhängige Benotung durchgeführt, wobei galt: 0=Kriterium nicht erfüllt, 1=Kriterium teilweise erfüllt, 2=Kriterium vollständig erfüllt.

In der ersten Phase der Beurteilung stand die Vereinbarkeit mit der Dateninnovationsstrategie 1.0 im Zentrum, wobei ein Produkt (der drei Bewertungen) von 0 einen Ausschluss zur Folge hatte.

1. Datentyp (primär oder sekundär).
2. Methoden (komplementäre, prädiktive, innovative).
3. Implementierbarkeit in die Produktion des BFS.

In der zweiten Phase der Bewertung wurden operative Aspekte der Pilotprojekte berücksichtigt, wobei der Mittelwert der Beurteilungen berechnet wurde ($\text{Score} = \text{Summe aller Bewertungen} / 10$), womit der Mittelwert (Score) zwischen 0 und 6 ist. Dieser Wert wurde aus Gründen der Interpretierbarkeit auf das Intervall $[1;6]$ projiziert ($(\text{Score} * 5/6) + 1$).

- 1) Zusammenstellung des Teams und deren (auto-evaluierter) Kompetenzen.
- 2) Automatisierung (Effizienzgewinn).
- 3) Transparenz (des Prozesses der zu den Resultaten führt).
- 4) Reproduzierbarkeit (in Bezug auf die angewandten Methoden).
- 5) Statistische Validierung (in Bezug auf die angewandten Methoden).
- 6) Dauer des Pilotprojekts.
- 7) Datenqualität.
- 8) Angebrachte Grösse und Dimension der Daten (n,p).
- 9) Mehrwert des Pilotprojekts.
- 10) Globale Relevanz (u.a. story telling, Kommunikation, Ethik).

Schlussendlich wurden die fünf Pilotprojektanträge mit den höchsten Werten zur Umsetzung vorgeschlagen.

Dieses Auswahlverfahren wird auch für die Auswahl weiterer zukünftiger Pilotprojekte angewandt. Falls das Pilotprojekt eine Einheit ausserhalb des BFS betrifft, bezieht sich insbesondere das Kriterium 3) auf diese Einheit. Dies setzt voraus, dass die externe Einheit einen nachvollziehbaren Prozessbeschreibung liefert (u.a. inklusive Beschreibung der zu verwendenden Daten

und deren Bereitstellung) und aufzeigt, wie das Pilotprojekt möglicherweise in ihre Produktion übernommen werden kann. Eine Vorabklärung mit dem BFS ist in diesen Fällen ratsam.

2.3 Kommunikation mit AG DS und GL

Die AG Data Science (AG DS) wird an ihren periodischen Sitzungen über den Fortschritt dieses Dokuments informiert. Die AG DS beurteilt dieses Dokument, bevor es der GL unterbreitet wird.

Die vormalige AG NDS wurde am 21.10.2019 aufgefordert, dieses Dokument bis zum 06.11.2019 zu beurteilen. Die Kommentare der AG NDS Mitglieder wurden in diesem Dokument so weit möglich berücksichtigt.

Die GL wurde im GL-Workshop vom 29.05.2019 mittels einem Reporting von METH (Jean-Pierre Renfer) über die „Governance“ und die möglichen Statuszustände der experimentellen Statistiken informiert. Letztere werden datiert und in die Zustände «im Gange», «angehalten», «abgeschlossen» und «in Produktion» unterschieden. Die GL entscheidet über die Initialisierung eines Pilotprojekts und über dessen Transfer in die Produktion.

2.4 Geltungsbereich

Das vorliegende Dokument wird periodisch überarbeitet und die aktuellste Version (Versionsnummer und Datum als Referenz) hat für alle Dateninnovationsprojekte des BFS Gültigkeit.

Die Pilotprojekte werden mittels der Version beurteilt, die für die Beurteilung des Pilotprojektsantrags angewandt wurde.

3 Vorgehen für einen möglichen Transfer in die Produktion

Für den Transfer in die Produktion wird zuerst die Freigabe zu einem HERMES-Projekt (resp. BFS-Standard Projekt) beantragt. Erfolgt diese Freigabe, wird der Transfer in die Produktion in der Regel nach BFS-Standards als Projekt abgewickelt.

Für Teiltätigkeiten, die zum Beispiel der Interpretation von Analysen dienen, kann ein erleichteter Transfer beantragt werden. Anpassungen der Methoden, die in die Produktion transferiert werden konnten, werden entsprechend der bereits bestehenden Qualitätskriterien und den relevanten zusätzlichen Beurteilungskriterien gehandhabt, und müssen somit das hier beschriebene Vorgehen nicht mehr durchlaufen.

Die Freigabe für das BFS-Standard Projekt wird über einen Beurteilungsprozess gesteuert, siehe 3.1, der inhaltlich mittels einheitlicher Kriterien durchgeführt wird, siehe 3.2.

Begleitend zum Transfer in die Produktion innerhalb des HERMES-Projekts, ist ein Update auf der Microsite «Experimentelle Statistiken», im Einklang mit dem GL-Reporting vom 29.05.2019, anzugehen.

3.1 Beurteilungsprozess

Der Beurteilungsprozess soll so schlank wie möglich gehalten werden. Der folgende Ablauf ist vorgesehen:

- 1) Anfrage durch den Projektleiter des Pilotprojekts mittels einem Transferantrag, der der Sektion statistische Methoden (METH) zur Beurteilung unterbreitet wird.
- 2) Beurteilung durch METH mittels einheitlicher Beurteilungskriterien und Weiterleitung an den/die fachlich zuständigen Abteilungscheff/s. Falls nötig und zielführend, kann für die Beurteilung ebenfalls eine externe Beraterung in Data Science Projekten hinzugezogen werden.
- 3) Beurteilung Abteilungscheff/s (Transferantrag mit Empfehlung METH).
- 4) GL-Reporting / GL-Antrag.

Der Entscheid zum Transfer in die Produktion wird von der GL getroffen. METH erteilt eine Empfehlung zuhanden des/der fachlich zuständigen Abteilungscheffs, in welcher die Stellungnahme von METH begründet wird. METH kann eine unvollständige Anfrage zurückweisen, um eine Vervollständigung durch den Projektleiter des Pilotprojekts zu verlangen.

3.2 Beurteilung

Transferantrag

Die Beurteilung basiert auf dem Transferantrag, der eine detaillierte Dokumentation umfasst. Der Transferantrag muss in Bezug zu den für das Pilotprojekt relevanten Beurteilungskriterien vollständig sein, wobei in begründeten Fällen einzelne Punkte differenziert werden können.

Wie bereits beim Auswahlverfahren für die Pilotprojekte der Dateninnovationsstrategie 1.0 wird die Beurteilung wiederum durch eine unabhängige Benotung durchgeführt, wobei 0 = Kriterium nicht erfüllt, 1 = Kriterium teilweise erfüllt, 2 = Kriterium vollständig erfüllt. Kann ein Kriterium nicht beurteilt werden, ist die Beurteilungsstufe NA („Not Available“).

1. Vorhandensein einer detaillierten Dokumentation, die die Vorgehensweise nachvollziehbar beschreibt?
2. Wurden die vorgegebenen Ziele in Bezug auf die Produktion, die im Pilotprojektbeschrieb definiert wurden, erreicht?
3. Mehrwert für die Produktion?
4. Güte der Validierung in Bezug auf die angewandten Methoden?
5. Übereinstimmung mit gängigen Kriterien des BFS in den entsprechenden Produktionsschritten?

Für die Empfehlung von METH können (falls zielführend) die zusätzlichen Beurteilungselemente im Anhang herangezogen werden. Zudem kann eine Evaluation der Pilotprojektantragskriterien, siehe 2.2, in die Empfehlungen einfließen.

Minimale Beurteilung

Für eine positive Empfehlung zum Transferantrag darf keines der Kriterien die Beurteilungsstufe NA oder 0 aufweisen.

Die Kriterien 2 und 3 dürfen eine Beurteilung 1 haben, da eine Teilerreichung bereits zu einem Gewinn für die Produktion führen könnte.

Hingegen müssen die Kriterien 1, 4 und 5 mit 2 beurteilt worden sein.

Gesamtbeurteilung

Die Gesamtbeurteilung dient der Übersicht über alle Kriterien. Sie entspricht dem mit 5 multiplizierten Mittelwert der Beurteilung der beurteilbaren Kriterien. Somit weist die Gesamtbeurteilung Werte zwischen 0 und 10 auf. Um die nicht beurteilbaren Kriterien zu berücksichtigen, wird eine untere und eine obere Grenze berechnet. Die untere Grenze entspricht der Summe der beurteilbaren Kriterien und die obere Grenze wird analog mit dem Wert 2 für nicht beurteilbare Kriterien berechnet. Somit liegt die Gesamtbeurteilung immer zwischen diesen beiden Grenzen (inklusive).

4 Anhang: Zusätzliche Beurteilungselemente

Die zusätzlichen Beurteilungselemente orientieren sich an generellen, übergreifenden Elementen, an Elementen betreffend die Entwicklung des Pilotprojekts, sowie an Elementen betreffend die Auswirkungen bei der Übernahme in die Produktion. Letztere sind in der Regel erst innerhalb des BFS-Standard Projekts zu bearbeiten, können aber ebenfalls in die Empfehlungen einfließen.

I) Übergreifende Elemente

1. Allgemeiner nachvollziehbarer Beschrieb des Pilotprojekts.
2. Ziele und Rahmenbedingungen des Pilotprojekts, sowie eine Beurteilung der Zielerfüllung sowie Beitrag zur Dateninnovationsstrategie.
3. Welches ist der Mehrwert in Bezug auf die Problemstellung?
4. Welches ist die globale Relevanz?
5. Wie wird die Referenzgrösse („Baseline“), der Ist-Zustand (vor dem Pilotprojekt), gemessen, mit welchem die Resultate des Pilotprojekts verglichen werden („Baseline“ Vergleich)?
6. „Lessons learned“ im Verlauf der Pilotprojektphase.
7. Selbständiges Arbeiten des Pilotprojektteams im Rahmen des «Coachings» («training on the job»)?
8. Automatisierung (Effizienzgewinn)?

II) Entwicklung des Pilotprojekts

9. Detaillierter Beschrieb der verwendeten Daten, Datenquellen und deren Qualität (z.B. Abdeckung, Messwertfehler, Bearbeitungsfehler), sowie deren Verfügbarkeit und Vergleichbarkeit/Stabilität (mittelfristig und langfristig) und Verknüpfbarkeit. Auswirkungen auf die Metadaten?
10. Detaillierter und reproduzierbarer Beschrieb des «Feature Engineering» (Variablenkonstruktion).
11. Detaillierter Beschrieb der angewandten Methoden und deren Stärken und Schwächen (z.B. Stabilität/Robustheit der Methoden, Effizienz der Methoden bezüglich Prädiktion/Prediction und Erklärung/Schätzung, Vorgehen zur Auswahl der allfälligen „Hyperparameter“). Gibt es Alternativen oder Verbesserungspotenzial?
12. Detaillierter und reproduzierbarer Beschrieb der Methoden zur Validierung der Resultate inklusive Gütemass(e) in Bezug auf die Prädiktion/Prediction (z.B. Verallgemeinerungsfehler, „generalisation (prediction) error“) und die Erklärung/Schätzung (z.B. Modellfehler und Modellpräzision, „goodness of fit“), Vorgehen bei deren Berechnung(en). Gibt es Alternativen oder Verbesserungspotenzial?
13. Vorgehensprozess für die Qualitätssicherung, das Monitoring der angewandten Methoden (inklusive Häufigkeit/Frequenz der Aktualisierung der Methoden, „recalibration“ und „model change management“) und das Monitoring der verwendeten Daten, wenn diese Methoden in der statistischen Produktion angewandt werden. Beschrieb der relevanten Qualitätsmerkmale des Produktionsprozesses, z.B. sind die Resultate vergleichbar über die Zeit?

14. Detaillierter Beschrieb der produzierten Daten/Resultate und deren Qualität (z.B. Abdeckung, Messwertfehler, Bearbeitungsfehler), sowie deren Verfügbarkeit und Vergleichbarkeit/Stabilität (mittelfristig und langfristig) und Verknüpfbarkeit. Auswirkungen auf die Metadaten?
15. Automatisierung (Effizienzgewinn) in Bezug auf die Entwicklung des Pilotprojekts?

III) Auswirkungen bei der Übernahme in die Produktion und wie damit umgegangen wird (BFS-Standard Projekt)

16. Allgemeine Begründung des Transfers in die Produktion.
17. Zeitpunkt des Transfers. Ist dieser realistisch, auch in Bezug auf die nötigen Ressourcen, und mit Revisionsvorhaben (falls vorhanden) abgestimmt?
18. Sind juristische Anpassungen nötig? Falls ja, Stellungnahme von RD.
19. Kann der Datenschutz gewährleistet werden? Stellungnahmen DSK und ISBO.
20. Sind Anpassungen des Produktionsprozesses nötig? Konzeptueller Beschrieb des Transferprozesses.
21. Ist die Methode skalierbar (auf grössere Datenmengen, respektive unter unterschiedlichen Rahmenbedingungen anwendbar)?
22. Sind die angewandten Methoden wiederverwendbar?
23. Einbettung der entsprechenden Methoden, Resultate und Daten in der Produktionsumgebung?
24. Automatisierung (Effizienzgewinn) in Bezug auf die Produktion?
25. Sind Anpassungen der bestehenden Produktions-Metadaten nötig? Falls ja, wie wird dies umgesetzt und kommuniziert?
26. Ist eine Revision bereits publizierter Resultate nötig?
27. Wie wird mit Zeitreihen umgegangen? Braucht es Anpassungen, respektive eine Retropolation?
28. Beschrieb der verwendeten Software/Pakete/Bibliotheken (inklusive Versionsnummern) und Abklärung, ob diese gegebenenfalls in der Produktion eingesetzt werden können (falls nicht über Schnittstellen möglich).
29. Auswirkungen auf die Infrastruktur/Softwarelandschaft? Falls ja, konzeptueller Beschrieb (Skizze) der Infrastrukturanpassungen (gilt auch für Schnittstellen).
30. Sind Anpassungen des Publikationsprozesses nötig? Falls ja, Stellungnahme von DIAM.
31. Kommunikation: wie ist diese im BFS und gegenüber Nutzern (u.a. FEDESTAT, REGIOSTAT, KORSTAT) sichergestellt und auf potenzielle Änderungen ausgerichtet?
32. Auswirkungen auf die nötigen Personalressourcen?
33. Besteht die Notwendigkeit einer zusätzlichen Ausbildung des Produktionsteams?
34. Finanzielle Auswirkungen?

Annexe N.3: Description des variables

	Variable	Nom détaillé	Type	Description
1	LOCALID	Local-ID	Character	Identifiant d'une entreprise au niveau local
2	ENTID	ENT-ID	Character	Identifiant d'une entreprise au niveau ENT
3	NAME	Nom de l'entreprise	Character	Nom de l'entreprise
4	ADDITIONALBUSINESSNAME	Nom supplémentaire	Character	Nom supplémentaire si disponible
5	SIPA_ACTIVE	Registre SIPA	Integer	Variable binaire pour savoir si l'entreprise est active dans le registre SIPA: 0=non-actif/1=actif
6	FORET_ACTIVE	Registre FORET	Integer	Variable binaire pour savoir si l'entreprise est active dans le registre FORET: 0=non-actif/1=actif
7	STRU_UNIV_ACTIVE	Univers STRU	Integer	Variable binaire pour savoir si l'entreprise est active dans l'univers STRU: 0=non-actif/1=actif
8	PL_UNIV_ACTIVE	Univers PL	Integer	Variable binaire pour savoir si l'entreprise est active dans l'univers profiling: 0=non-actif/1=actif
9	PL_UNIV_DESCR	Description univers PL	Character	Description d'activité dans l'univers profiling
10	MAS_UNIV_ACTIVE	Univers MAS	Integer	Variable binaire pour savoir si l'entreprise est active dans l'univers MAS: 0=non-actif/1=actif
11	KS_UNIV_ACTIVE	Univers KS	Integer	Variable binaire pour savoir si l'entreprise est active dans l'univers KS: 0=non-actif/1=actif
12	SOMEDA_UNIV_ACTIVE	Univers Someda	Integer	Variable binaire pour savoir si l'entreprise est active dans l'univers SOMEDA: 0=non-actif/1=actif
13	SPITEX_UNIV_ACTIVE	Univers Spitex	Integer	Variable binaire pour savoir si l'entreprise est active dans l'univers SPITEX: 0=non-actif/1=actif
14	HESTA_UNIV_ACTIVE	Univers Hesta	Integer	Variable binaire pour savoir si l'entreprise est active dans l'univers HESTA: 0=non-actif/1=actif
15	NOGACODE	Code Noga	Character	Code Noga attribué
16	NOGACODE_TX	Code Noga texte	Character	Description du Code Noga
17	TYPE	Type	Integer	Catégorie du type d'entreprise
18	TYPE_TX	Type texte	Character	Description de la catégorie du type d'entreprise

19	EMPL_MEN	Emploi homme	Numeric	Nombre d'employé de sexe masculin dans une entreprise
20	EMPL_WOMEN	Emploi femme	Numeric	Nombre d'employé de sexe féminin dans une entreprise
21	EMPL_TOTAL	Emploi total	Numeric	Nombre total d'employé dans une entreprise
22	FTE_RECALC	Full time employment recalculé	Numeric	Nombre d'employé recalculé à temps plein dans une entreprise
23	FTE_MEN	Full time employment homme	Numeric	Nombre d'employé de sexe masculin à temps plein dans une entreprise
24	FTE_WOMEN	Full time employment femme	Numeric	Nombre d'employé de sexe féminin à temps plein dans une entreprise
25	FTE_TOTAL	Full time employment total	Numeric	Nombre total d'employé à temps plein dans une entreprise
26	STREET	Rue	Character	Rue
27	TOWN	Ville	Character	Ville
28	CANTON	Canton	Integer	Catégorie du canton (1-26)
29	CANTON_TX	Canton Texte	Character	Nom du canton
30	LANGUAGE	Langue	Integer	Catégorie de langue (1-4)
31	LANGUAGE_TX	Langue texte	Character	Nom de la langue
32	COORDE	Coordonnée est	Integer	Coordonnée est
33	COORDN	Coordonnée nord	Integer	Coordonnée nord
34	EGID	Identificateur de bâtiment	Integer	Identifiant de l'identificateur de bâtiment
35	LEG_AVS_ACTIVE	Registre AVS	Integer	Variable binaire pour savoir si l'entreprise est active dans le registre AVS: 0=non-actif/1=actif
36	LEG_AVS_DESCR	Description registre AVS	Character	Description d'activité dans le registre AVS
37	LEG_TVA_ACTIVE	Registre TVA	Integer	Variable binaire pour savoir si l'entreprise est active dans le registre TVA: 0=non-actif/1=actif
38	LEG_TVA_DESCR	Description registre TVA	Character	Description d'activité dans le registre TVA
39	LEG_RC_ACTIVE	Registre RC	Integer	Variable binaire pour savoir si l'entreprise est active dans le registre RC: 0=non-actif/1=actif

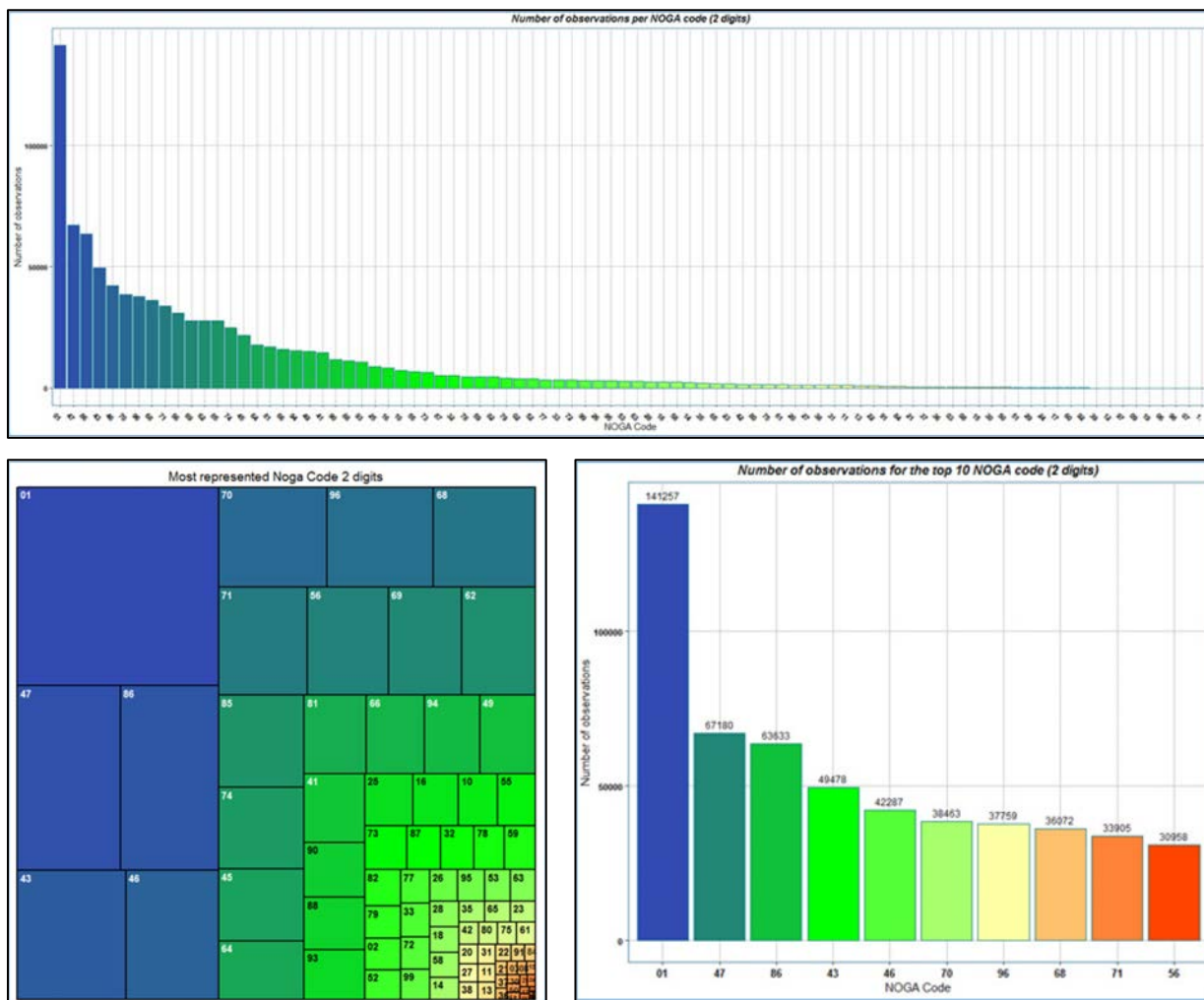
40	LEG_RC_DESCR	Description registre RC	Character	Description d'activité dans le registre RC
41	LEG_MEDREG_ACTIVE	Registre MEDREG	Integer	Variable binaire pour savoir si l'entreprise est active dans le registre MEDREG: 0=non-actif/1=actif
42	LEG_MEDREG_DESCR	Description registre MEDREG	Character	Description d'activité dans le registre MEDREG
43	LEG_AVANDORNOT_ACTIVE	Registre avocat et notaire	Integer	Variable binaire pour savoir si l'entreprise est active dans le registre avocat et notaire: 0=non-actif/1=actif
44	LEG_SIPA_ACTIVE	Registre SIPA	Integer	Variable binaire pour savoir si l'entreprise est active dans le registre SIPA: 0=non-actif/1=actif
45	LEG_DOUANE_ACTIVE	Registre douane	Integer	Variable binaire pour savoir si l'entreprise est active dans le registre douane: 0=non-actif/1=actif
46	LEG_LEGAL_FORM	Raison sociale	Integer	Catégorie de forme juridique
47	LEG_LEGAL_FORM_TX	Raison sociale texte	Character	Description de la forme juridique
48	LEG_BIRTHDATE	Date de fondation	Integer	Date de fondation de l'entreprise (création de 3 colonnes dans le R)
49	ENT_NOGACODE	Code Noga sous l'ENT_ID	Integer	Numéro Code Noga
50	ENT_NOGACODE_TX	Texte Code Noga sous l'ENT_ID	Character	Description Code Noga
51	ENT_TURNOVER	Chiffre d'affaires CHF	Numeric	Chiffre d'affaires annuel de l'entreprise en CHF
52	ENT_EXPORTTOTALAMOUNT	Valeurs des exportations en CHF	Numeric	Montant des exportations en CHF
53	ENT_IMPORTTOTALAMOUNT	Valeurs des importations en CHF	Numeric	Montant des importations CHF
54	ENT_MAINEXPORTED PRODUCT_TX	Texte produit principal exporté	Character	Produit principal exporté
55	ENT_MAINIMPORTED PRODUCT_TX	Texte produit principal importé	Character	Produit principal importé

56	ENT_MONO_MULTI_ ADMIN_LOC	Mono- ou multi établissement	Character	Entreprise de catégorie mono- ou multi-établissement
----	------------------------------	------------------------------	-----------	--

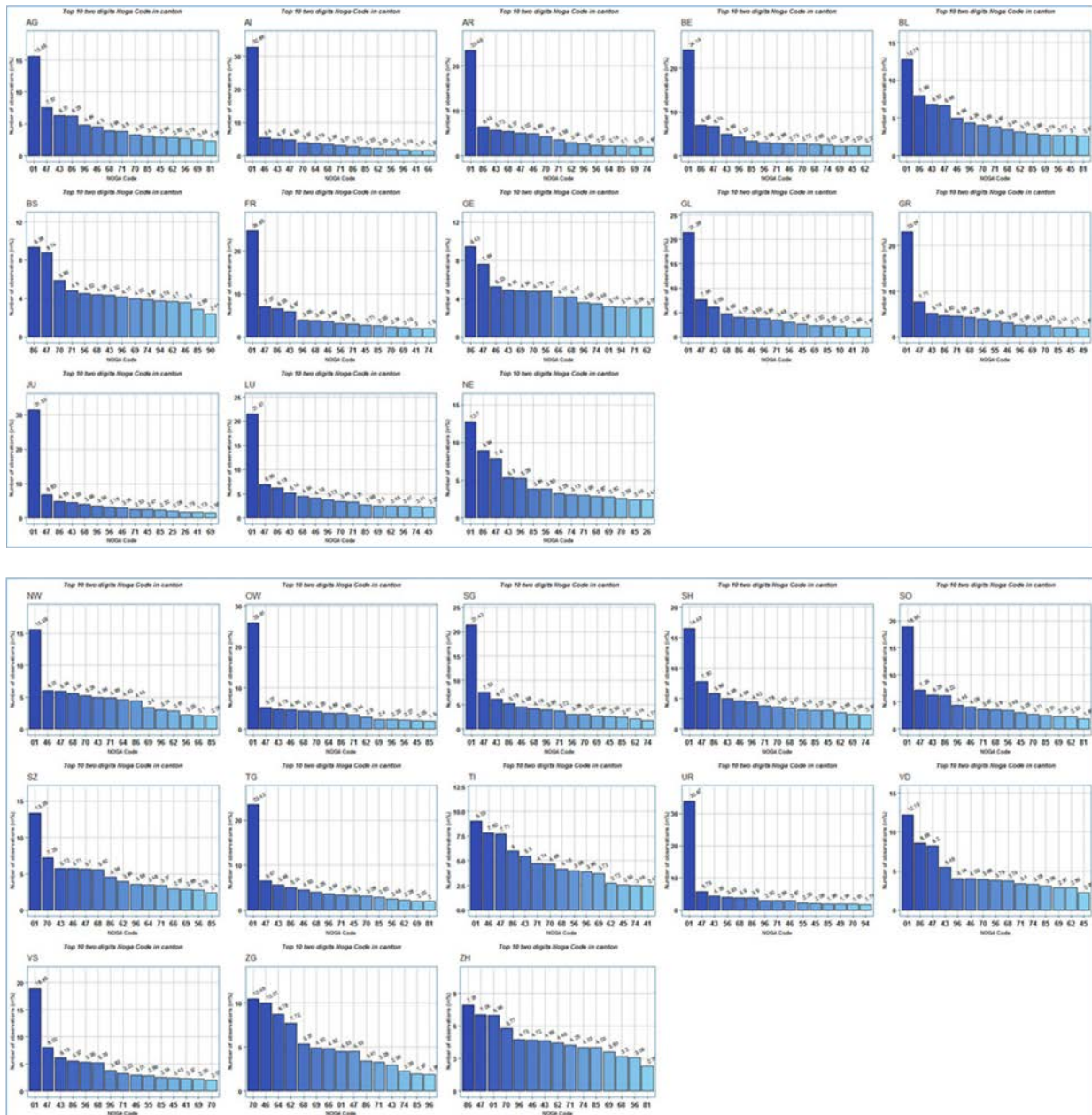
Annexe N.4: Exploration des données

La partie « Explore Data » nous permet d'avoir une première impression de la totalité du jeu de données. Les premières vues concernent la répartition des codes NOGA.

Le code NOGA à deux chiffres le plus représenté avec environ 15% des entreprises enregistrées est celui du 01 "**Culture et production animale, chasse et services annexes**", suivi de 47 "**Commerce de détail, à l'exception des automobiles et des motocycles**" et en troisième position 86 "**Activités pour la santé humaine**".



Ci-dessous nous avons la répartition des codes NOGA à deux chiffres conditionnés par canton. Nous remarquons que pour 4 cantons sur 26 ce n'est pas le code NOGA 01 qui est le plus représenté au sein du canton mais le code NOGA 86 ou encore 70 (Activités des sièges sociaux; conseil de gestion) pour le cas du canton de Zoug. Ceci peut simplement être expliqué par le fait que ce sont des cantons urbanisés ou dans le cas de Zoug avec des avantages fiscaux.



Voici la comparaison de la distribution des genres humains par codes NOGA. Dans le tableau ci-dessus, nous voyons que le code NOGA 39 ("Dépollution et autres services de gestion des déchets") a en moyenne plus d'hommes que de femmes mais cela n'est représenté que par 38 entreprises alors que le code NOGA 96 ("Autres services personnels") est représenté plus de 28'000 fois et a en moyenne plus de femmes. Cependant, représenté plus de 39'000 fois, le code NOGA 43 ("Travaux de construction spécialisés") emploie en moyenne 87.20% du temps des hommes.

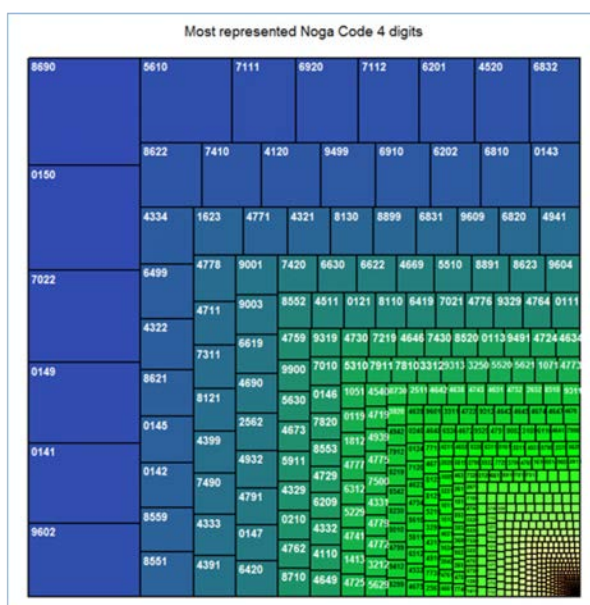
Les codes NOGA où les deux genres sont équilibrés sont:

- 58: Édition;
- 13: Fabrication de textiles;
- 93: Activités sportives, récréatives et de loisirs;
- 68: Activités immobilières;
- 73: Publicité et études de marché.

Noga Code two digits	Mean men in %	Mean women in %	Number of companies	Noga Code two digits	Mean men in %	Mean women in %	Number of companies	Noga Code two digits	Mean men in %	Mean women in %	Number of companies
06	100.00	0.00	1	95	74.68	25.32	2231	90	55.20	44.80	8096
98	100.00	0.00	1	50	74.52	25.48	199	92	54.73	45.28	70
39	92.78	7.22	38	27	73.86	26.14	879	74	54.39	45.61	17644
42	91.67	8.33	940	77	72.04	27.96	1954	51	52.70	47.30	228
02	90.31	9.69	1404	78	71.83	28.17	3549	93	52.29	47.71	6564
41	89.11	10.89	8627	17	71.82	28.18	179	65	51.47	48.53	1177
37	87.50	12.50	438	12	71.33	28.67	13	58	50.57	49.44	1496
43	87.20	12.80	39058	52	71.15	28.85	2348	13	49.64	50.36	777
08	87.14	12.86	329	22	70.22	29.78	760	68	49.26	50.74	19283
30	86.52	13.48	253	20	70.09	29.91	824	73	49.03	50.97	4344
24	85.40	14.60	272	01	69.68	30.32	143	56	47.83	52.16	23592
16	85.31	14.69	7046	11	69.54	30.46	648	84	46.59	53.41	510
33	84.99	15.01	2767	71	69.35	30.65	25389	82	46.14	53.86	3007
29	84.21	15.79	234	63	68.74	31.26	1341	94	45.31	54.69	6465
49	83.42	16.57	11067	01	63.85	36.15	51016	81	44.96	55.05	12887
38	83.27	16.73	848	46	63.54	36.46	26155	69	44.91	55.09	21804
28	82.76	17.24	2040	66	62.42	37.58	11068	55	44.33	55.68	5047
45	81.96	18.03	16819	18	61.96	38.04	2063	85	43.27	56.73	18540
23	81.87	18.14	1413	59	61.88	38.12	3156	15	43.17	56.82	309
09	79.99	19.99	15	26	61.28	38.72	2195	79	37.22	62.78	2752
35	79.93	20.07	1023	72	60.15	39.85	1862	91	35.75	64.25	513
25	79.77	20.23	7254	64	59.95	40.05	9166	47	31.96	68.05	48736
19	79.46	20.54	7	60	59.43	40.57	147	88	28.19	71.81	7635
31	78.72	21.28	893	70	58.70	41.30	24299	86	23.66	76.34	51189
62	78.52	21.50	17995	53	57.82	42.18	2638	14	23.65	76.35	1531
36	78.26	21.74	277	21	57.19	42.81	314	87	23.46	76.54	3544
80	78.25	21.75	948	32	57.16	42.84	3722	75	20.11	79.89	1270
61	74.81	25.19	1137	10	55.97	44.03	4373	96	18.91	81.09	28090

Classement des codes NOGA à 4 digits les **plus** représentés dans le jeu de données:

1. 8690: **Autres activités pour la santé humaine**
2. 0150: **Culture et élevage associés**
3. 7022: **Conseil pour les affaires et autres conseils de gestion**



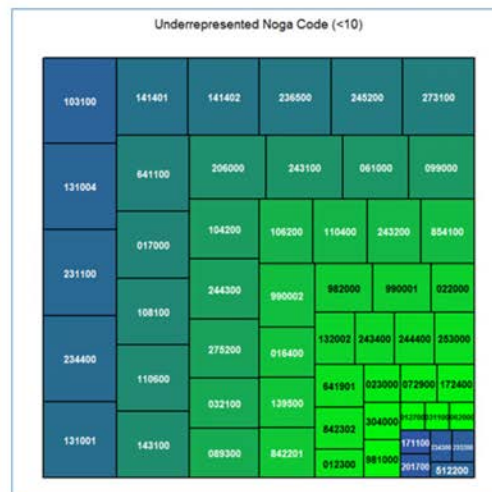
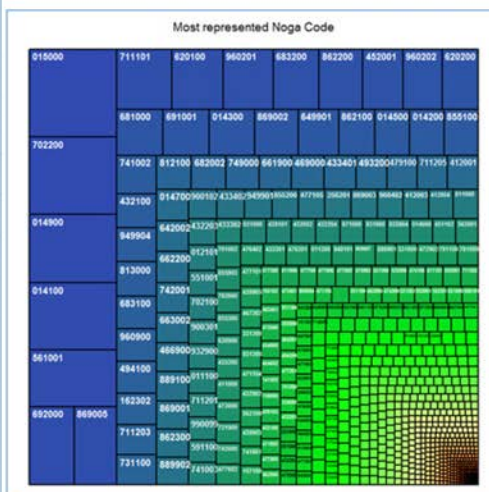
Classement des codes NOGA à 6 digits les **plus** représentés dans le jeu de données:

- 015000: Culture et élevage associés;
- 702200: Conseil pour les affaires et autres conseils de gestion;
- 014900: Élevage d'autres animaux

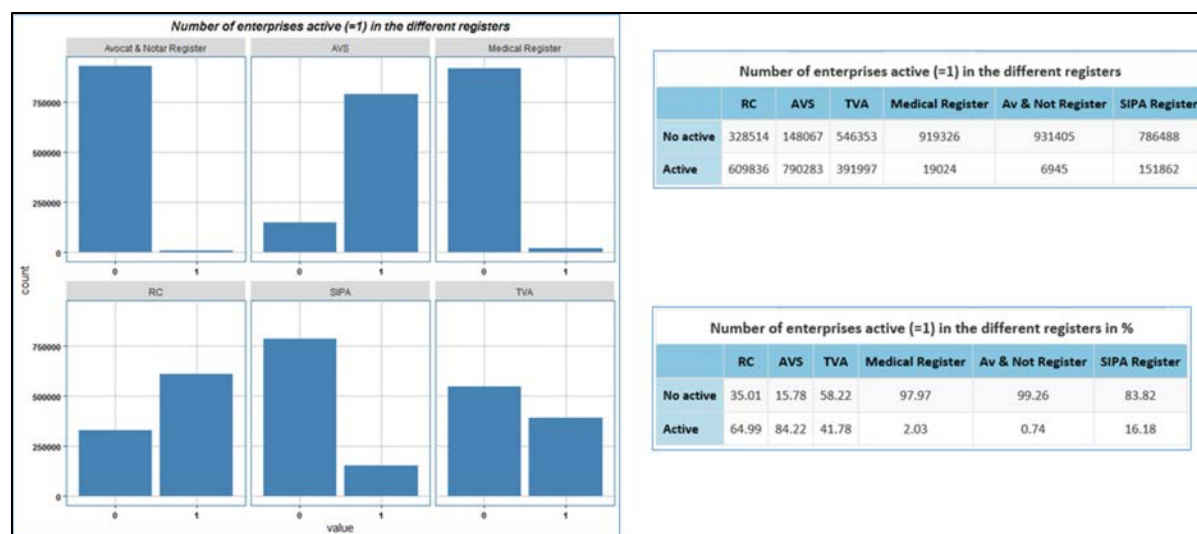
Les codes NOGA à 6 digits les **moins** représentés dans le jeu de données:

- 512200: Transports spatiaux;
- 234300: Fabrication d'isolateurs et pièces isolantes en céramique;
- 235200: Fabrication de chaux et plâtre

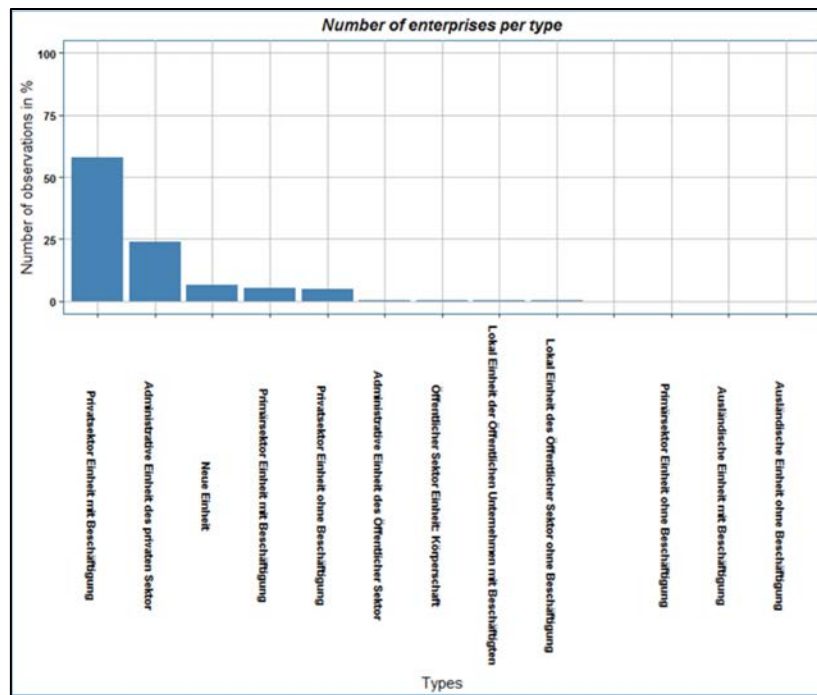
Most represented Noga Code		
Noga Code	Number of companies	% of companies
015000	38991	3.94
702200	32400	3.45
014900	28408	3.03
014300	28207	3.01
563001	23677	2.52
692000	16728	1.78
869005	16129	1.72
711101	15869	1.69
620900	14913	1.59
960201	14104	1.55
681200	13136	1.42
862200	12626	1.35
452001	11626	1.24
960302	11390	1.21
620200	10109	1.10
681000	10080	1.10
693001	10054	1.09
014300	10058	1.07
869002	9653	1.03
649901	8961	0.95
862300	7962	0.85
014500	7795	0.83
014200	7640	0.81
805300	7496	0.80
743002	7323	0.78
432100	6542	0.70
949904	6497	0.69
813000	6435	0.69



Ci-dessous, nous pouvons voir dans quel registre il y'a le plus d'entreprises actives. Les registres AVS et RC ont respectivement 84.22% et 65% d'entreprises actives dans leur base de données. C'est ainsi dans ces registres que nous allons certainement retrouver le plus d'informations sur les activités des entreprises.



Le graphique ci-dessous, nous indique que plus de la moitié des entreprises sont de type "Unité du secteur privé avec occupation" suivi du type "Unité administrative du secteur privé". Ces deux types représentent à eux seul le 80% des types d'entreprises dans notre jeu de données. La troisième catégorie la plus importante, "Neue Einheit/Nouvelle Entité", est attribuée à une unité locale qui n'était pas (ou plus) statistiquement active, et qui se réactive statistiquement (via un ajout d'un univers dans le REE) mais qui n'a encore jamais eu de participation pour ces univers. En théorie, cela revient à une toute nouvelle entreprise (qui vient d'être créée dans le REE) ou à une entreprise existante qui n'a jamais participé à aucune enquête.



Annexe N.5: R packages pour la détection de la langue

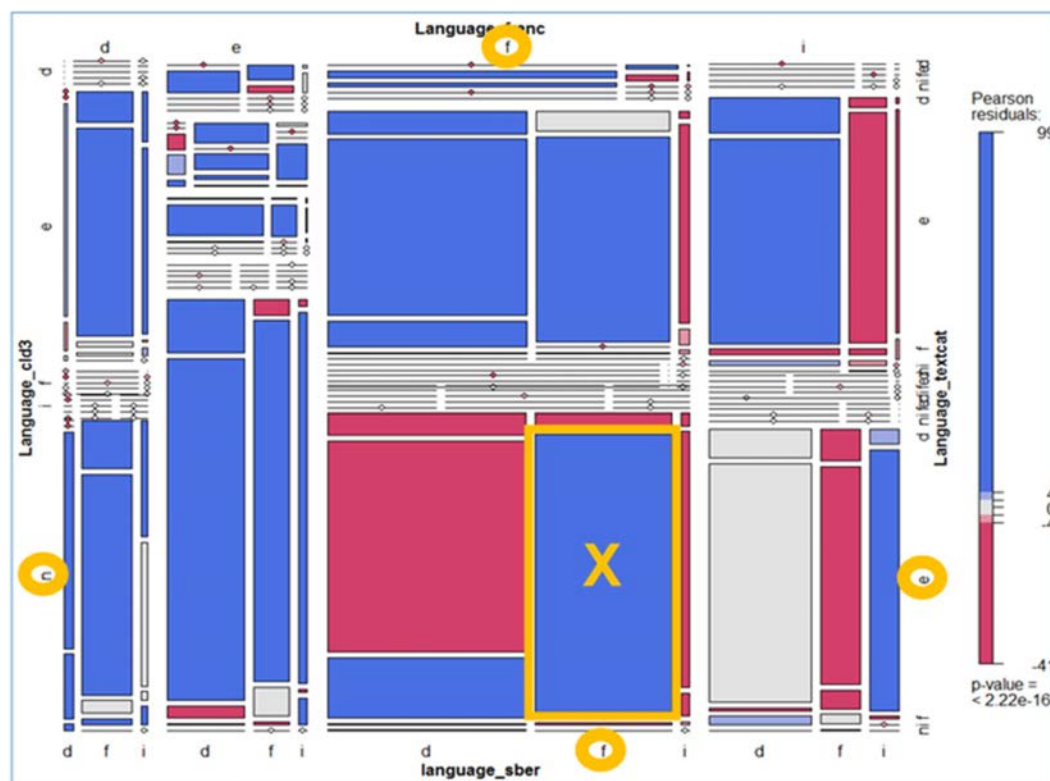
Packages	Created by	Method used	Number of languages
Package «textcat» Version: 1.0-6 Year: 2018	Kurt Hornik, Johannes Rauch, Christian Buchta, Ingo Feinerer	<i>“This package uses two steps. In a first time the package collects, training corpus of texts all written in the same known language and from there builds language profiles. The algorithm classifies from the most to the least frequent n-grams and only keeps the most frequent ones. In a second time, from the unknown language of a given text, it computes again a document profile and classifies it according to the language profile performed in the first step. It minimises the “out-of-place”¹ distance between the n-gram profiles of both texts and categorizes as the most similar.</i> References, CRAN: https://cran.r-project.org/web/packages/textcat/textcat.pdf References, Researchgate: https://www.researchgate.net/publication/282255965_The_textcat_Package_for_n-Gram_Based_Text_Categorization_in_R ¹ https://github.com/halfvim/N-Gram/wiki/Out-of-place-Distance-Measure	74 languages
Package «cld3» Version: 1.3 Year: 2020	Jeroen Ooms, Google Inc	Google's Compact Language Detector 3: <i>“CLD3 is a neural network model for language identification. The model averages the embeddings corresponding to each ngram type according to the fractions, and the averaged embeddings are concatenated to produce the embedding layer. The remaining components of the network are a hidden (Rectified linear) layer and a softmax layer. To get a language prediction for the input text, the model performs a forward pass through the network.”</i> References, Github: https://github.com/google/cld3#readme	47 languages
Package «cld2» Version: 1.2 Year: 2018	Jeroen Ooms, Dirk Sites	Google's Compact Language Detector 2: <i>« For mixed-language input, CLD2 returns the top three languages found and their approximate percentages of the total text bytes using the Naïve Bayesian classifier.</i> References, Github: https://github.com/cld2owners/cld2#readme	83 languages
Package «franc» Version: 1.1.2 Year: 2019	Gabor Csardi, Titus Wormer, Maciej Ceglowski, Jacob R. Rideout, Kent S. Johnson	<i>Possibility to reduce the scope to 3 languages: German / French / Italian</i> References, Github: https://github.com/gaborcsardi/franc#readme	310 languages

Autre package utilisé :

Rcpp	plyr	shiny	openxlsx
RSQLite	purrr	xgboost	tidyverse
arules	readxl	shinymanager	jsonlite
readr	shiny	flexdashboard	glue
cld2	stringi	shinythemes	shiny.i18n
cld3	stringr	shinyWidgets	shinydashboardPlus
dplyr	tokenizers	shinydashboard	plotly
franc	text2vec	shinycssloaders	arules
kableExtra	textcat	shinyjs	FactoMineR
knitr	tm	reactable	ggplot2
Matrix	writexl	readxl	vcd
kableExtra			

Annexe N.6: Présentation graphique mosaïque

Le graphique ci-dessous représente la meilleure combinaison pour cette technique d'analyse. Le rectangle le plus volumineux dans un graphique représente la combinaison entre variables la plus récurrente. Il faut détecter le plus grand rectangle avec la plus haute valeur de Pearson residuals (cette valeur montre la force de la relation ainsi que la direction des associations entre les variables, la valeur représente la valeur absolue des résidus) et au moins deux langues similaires comme ci-dessous. Le graphique représente visuellement la première ligne mosaic1 du tableau en page 8.



Annexe N.7: Présentation « word embedding visualisation »

