Nº 11-522-x au catalogue
ISSN : 1709822-X

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Modèles de forêts aléatoires, une proposition pour l'analyse de stratégies de vérification sélective

par Fabiana Rocci, Roberta Varriale et Salvatore Coppola

Date de diffusion : le 15 octobre 2021

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Modèles de forêts aléatoires, une proposition pour l'analyse de stratégies de vérification sélective

Fabiana Rocci, Roberta Varriale et Salvatore Coppola[1]

Résumé

L'Institut national italien de statistique (Istat) a lancé un nouveau projet relatif aux processus statistiques à court terme, afin de respecter la future nouvelle réglementation européenne visant à diffuser des estimations plus rapidement. L'évaluation et l'analyse du processus d'enquête actuel de l'enquête à court terme sur le chiffre d'affaires dans les services (FAS) visent à relever la façon dont les meilleures caractéristiques des méthodes et pratiques actuelles peuvent être exploitées pour concevoir un processus plus « efficace ». Le projet devrait, en particulier, fournir des méthodes permettant d'appliquer en général d'importantes économies d'échelle, de portée et de connaissances au contexte de production des enquêtes à court terme, qui ont généralement recours à un nombre restreint de ressources. L'analyse du processus TEL QUEL a révélé que l'enquête FAS entraînait des coûts substantiels de vérification et d'imputation, en particulier du fait de l'importance du suivi et de la vérification interactive utilisés pour tous les types d'erreurs détectées. Dans cette optique, nous avons essayé d'exploiter les leçons retenues en participant au Groupe de haut niveau sur la modernisation des statistiques officielles (GHN-MSO, CEE-ONU) sur l'utilisation de l'apprentissage automatique dans les statistiques officielles. Dans cette étude, nous présentons une première expérience utilisant les modèles de forêt aléatoire pour : (i) prédire les unités représentant des données « douteuses », (ii) évaluer l'utilisation du potentiel de prédiction sur de nouvelles données et (iii) explorer des données afin de relever des règles et des tendances cachées. Nous nous concentrons en particulier sur le recours à la modélisation par forêt aléatoire pour comparer certaines autres méthodes en matière d'efficacité de la prédiction d'erreurs et pour traiter des aspects principaux de la nouvelle conception du schéma de vérification et d'imputation.

1. Introduction

Les méthodes d'apprentissage automatique (AA) sont considérées comme ayant un grand potentiel également dans le contexte des organismes nationaux officiels de statistique (ONS). Pour répondre à ce besoin, le Groupe de haut niveau sur la modernisation des statistiques officielles (GHN-MSO) a lancé un projet d'apprentissage automatique en 2019. Le projet visait à montrer la valeur ajoutée de l'apprentissage automatique, à savoir s'il permettait de produire des données plus pertinentes, actuelles, exactes et fiables de façon efficace. Pour les ONS, l'un des principaux enjeux consiste à savoir comment exploiter l'AA pour automatiser certains processus ou du moins certaines phases de processus ou encore aider les humains à les exécuter.

Lancé en avril 2019, le projet réunissait alors 23 participants issus de 13 organismes. Il est ensuite passé à 120 membres de 23 pays, 31 organismes nationaux et 4 organisations internationales. Il a duré deux années, pendant lesquelles de nombreuses idées ont été évaluées et de nouvelles idées ont vu le jour. Des rapports et d'autres documents ont été publiés sur 19 études pilotes, les premiers développements de l'utilisation de l'apprentissage automatique aux fins de vérification des données, et un pipeline générique de production de statistiques officielles à l'aide de données satellitaires et de l'apprentissage automatique. Dans cet article, nous aimerions présenter les premières idées et les développements qui ont suivi pour analyser la mesure dans laquelle les algorithmes d'apprentissage automatique peuvent servir à améliorer efficacement la vérification et l'imputation. Nous étudions plus particulièrement l'étape de vérification, qui est souhaitée dans toutes les méthodes et permet de déceler les erreurs non dues à l'échantillonnage dans les données.

Dans un premier temps, les principaux aspects à évaluer sont liés à la capacité potentielle de l'apprentissage automatique à tirer des leçons des données, à prédire automatiquement les erreurs potentielles, et à traiter une très

[1]Fabiana Rocci, Istat, Département de méthodologie, Rome, Italie; Roberta Varriale, Département de méthodologie, Rome, Italie; Salvatore Coppola, Istat, Département de la production régionale, Naples, Italie

grande quantité de données, pour lesquelles il peut être très difficile de rechercher une tendance sous-jacente inconnue. À la suite d'expériences, un autre aspect inattendu nous est apparu : le gain potentiel d'information sur la façon de remanier les processus statistiques en choisissant parmi différentes stratégies.

Nous examinerons le remaniement des processus statistiques de vérification et d'imputation (V et I) dans les statistiques à court terme (SCT) à l'Institut national italien de statistique (Istat), qui vise à satisfaire aux nouvelles exigences du règlement sur les SCT d'Eurostat, qui impliquent généralement de très fortes contraintes de temps pour les processus. À partir des toutes premières idées d'identification et de prédiction des « données erronées douteuses » à l'aide de l'apprentissage automatique, nous présenterons ce que nous avons découvert en cours de route et quelques pistes que nous essayons de proposer sur la façon d'effectuer la V&I en utilisant autrement les méthodes d'apprentissage automatique.

Après analyse des procédures actuelles de vérification des SCT, nous avons constaté qu'elles sont fortement fondées sur l'identification et le traitement des erreurs influentes, au moyen de plusieurs méthodes de vérification sélective. Nous avons ensuite exploité des données historiques et mis à l'essai une nouvelle méthode sélective, basée sur un modèle à classes latentes (SeleMix). En appliquant, dans une nouvelle expérience, l'apprentissage automatique aux fins de prédiction des erreurs influentes, nous avons découvert que l'identification de tendances cachées peut être un gain majeur apporté par ces méthodes. Dans le présent article, nous montrerons comment nous essayons d'éclairer les domaines d'amélioration possibles du processus d'enquête.

2. Le projet de remaniement du procédé de V et I de l'enquête FAS

Un nouveau Règlement relatif aux statistiques européennes d'entreprises d'Eurostat visera à obtenir de meilleurs résultats dans plusieurs types de publications, notamment pour ce qui est du niveau de détail, des normes et de la fréquence des résultats publiés. Dans ce contexte, le règlement relatif à l'enquête à court terme sur le chiffre d'affaires dans les services (FAS) devrait être modifié, parce qu'il est actuellement rédigé en vertu du règlement de l'UE (CE) no 1165/98 du Conseil et de ses amendements. La FAS diffuse des indices trimestriels du chiffre d'affaires – qui représente la variable cible – par entreprise du secteur des services aux prix courants; ils indiquent les variations de la variable cible, mesurées selon la quantité des ventes, comparativement à une année de référence fixe (année de référence), pour l'activité économique représentée par le secteur G, H, I, J, M, N de la Nomenclature statistique des activités économiques dans les communautés européennes (NACE) Rév.2. Le principal changement de la FAS en cours d'évaluation concerne le passage d'une diffusion trimestrielle à une diffusion mensuelle des indicateurs, ce qui modifierait considérablement le processus.

L'enquête FAS est organisée en sous-processus, chacun conçu en fonction d'un plan de V et I commun, mais caractérisé par des caractéristiques différentes. Nous avons sélectionné pour le premier essai le sous-processus concernant la division 46 de la NACE (G 46 – Commerce de gros, à l'exception des automobiles et des motocycles), car elle comprend déjà une procédure de vérification sélective. Notre objectif est d'évaluer l'efficacité de la stratégie actuelle, de proposer éventuellement certains changements, et d'utiliser les résultats pour introduire une phase de vérification sélective dans les autres sous-processus. La nouvelle stratégie doit être plus efficace, tant sur le plan des ressources humaines que du temps.

La division 46 de la NACE dans la FAS recueille des renseignements sur le chiffre d'affaires dans les services des entreprises classées par la NACE dans le commerce de gros, à l'exception des automobiles et des motocycles. L'échantillon de l'enquête est un panel d'entreprises, qu'on sélectionne à l'année de référence selon des critères d'échantillonnage par quotas, afin d'atteindre les 70 % du chiffre d'affaires total mesuré par le registre des entreprises italiennes (ASIA). L'échantillon de l'enquête compte ainsi environ 4 500 entreprises.

Le modèle TEL QUEL pour le processus de V et I a été représenté en macrophases par rapport au modèle générique d'édition de données statistiques (GSDEM; CEE-ONU, 2019), ce qui signifie qu'on détermine d'abord le domaine et les erreurs systématiques, au moyen de règles de vérification déterministes fondées sur la variation au cours de l'année de la variable cible. Ensuite, une méthode de vérification sélective permettant de repérer les erreurs influentes est exécutée.

Presque tous les critères servant à détecter tous les types d'erreurs sont fondés sur le profil longitudinal des entreprises elles-mêmes, qui se base sur la comparaison entre les données historiques des mêmes unités statistiques. Pour chaque type d'erreur (de domaine, systématique et influente), le traitement est exécuté par des méthodes interactives.

Le principal changement prévu est le lancement de la diffusion mensuelle des données. À cette fin, un nouveau projet a commencé à concevoir un nouveau procédé de V et I, qui maintiendrait un certain niveau de qualité en réduisant l'intervention humaine.

Étant donné que l'analyse des procédures actuelles a révélé que l'enquête FAS entraîne des coûts importants de V et I, dans lesquels la vérification sélective représente une grande part, les résultats de notre étude devraient orienter l'intervention humaine au cours de certaines phases du processus, et ainsi réduire les coûts, tout en respectant les exigences de rapidité et en augmentant l'efficacité. C'est pourquoi notre étude réalise une première expérience d'utilisation de modèles de forêt aléatoire pour à la fois prédire les unités représentant des données « douteuses », évaluer l'utilisation du potentiel de prédiction sur de nouvelles données, et explorer des données afin de relever des règles et des tendances cachées.

La stratégie du projet a commencé par l'analyse des données de la stratégie actuelle de V et I.
Les étapes programmées sont les suivantes : (i) mettre à l'essai une autre méthode pour évaluer la possibilité de déterminer les unités les plus « douteuses » qu'il faut traiter de façon interactive en maintenant un degré donné de qualité dans les estimations finales; (ii) appliquer l'idée lancée entre-temps par l'un des groupes de travail du Groupe de haut niveau sur la modernisation des statistiques officielles (GHN-MSO, CEE-ONU) sur l'utilisation de l'apprentissage automatique dans les statistiques officielles; (iii) utiliser l'apprentissage automatique pour prédire des données « douteuses » afin d'évaluer la capacité potentielle réelle de l'apprentissage automatique comme outil de prédiction dans la phase de V et I.

La méthode de vérification actuelle (méthode I) de la division 46 de la NACE dans la FAS se fonde sur les résultats de deux procédures de vérification : la procédure A et la procédure B. L'idée principale des deux est de définir une limite d'acceptation qui varie en fonction de la taille d'une unité en termes de quantité de la variable cible. En particulier, les deux méthodes élaborent une valeur transformée pour chaque enregistrement qui comprend un facteur de variation en pourcentage et un facteur de taille qui est ajusté par les paramètres de la méthode. La procédure A, fondée sur la méthode de Hidiroglou-Berthelot, identifie habituellement beaucoup plus d'unités présentant une erreur influente que la procédure B. La classification finale d'une unité en erreurs influentes potentielles/non potentielles est obtenue comme étant l'intersection entre la procédure A et la procédure B : c.-à-d. Méthode I $\equiv$ Procédure A $\cap$ Procédure B. En moyenne, le nombre d'erreurs influentes détectées et traitées chaque trimestre est d'environ 9 % du panel total de l'enquête. Le tableau 2 présente les résultats de la méthode I pour le nombre total d'enregistrements recueillis en 2018.

Tableau 2.
Distribution du nombre d'enregistrements influents (FAS NACE 46.) – année 2018

Enregistrements influents	fréquence	pourcentage
Oui	1 507	9,4
Non	14 466	90,6
Total	15 973	

3. Essai d'une autre méthode de vérification sélective basée sur le progiciel SeleMix

Nous mettons en œuvre la méthode II de vérification sélective qu'il faut tester dans le progiciel SeleMix (Guarnera et Buglielli, 2013; Buglielli et Guarnera, 2016) disponible dans R. Cette méthode est basée sur un modèle à classes latentes, tirant parti d'une spécification probabiliste des données réelles et du mécanisme d'erreur. Plus précisément, on suppose un modèle gaussien pour les données réelles et un mécanisme d'erreur « intermittent », de sorte qu'une proportion des données est contaminée par une erreur additive gaussienne (Di Zio et Guarnera, 2013). Les observations sont classées par ordre de priorité en fonction des valeurs d'une fonction de score qui exprime l'effet de leur erreur potentielle sur les estimations d'intérêt (Latouche et Berthelot, 1992), ce qui donne l'ordre des unités en ce qui

concerne leur risque d'être de « vraies » erreurs influentes. Toutes les unités au-dessus d'un seuil donné sont sélectionnées pour être traitées de manière interactive puisqu'elles représentent potentiellement les observations présentant des erreurs importantes. Le modèle utilisé par SeleMix pour la FAS – NACE 26 est conçu comme suit :

- Variable cible : Chiffre d'affaires au trimestre T
- Variable-covariable : Chiffre d'affaires au trimestre T-4

Le modèle est exécuté sur chaque trimestre pour chaque strate (donnée par groupe d'activités de la NACE à 3 chiffres, par classe de taille de l'entreprise).

Soulignons que dans notre étude expérimentale, nous ne pouvions utiliser que les corrections sur les unités sélectionnées par la méthode I, car il est impossible d'avoir les valeurs corrigées pour les unités sélectionnées par la méthode II et non sélectionnées par la méthode I. Cela explique que pour évaluer l'efficacité des deux méthodes de vérification sélective, nous avons utilisé le nombre absolu d'unités sélectionnées par les deux méthodes (unités se chevauchant) comme erreurs influentes potentielles, et le pourcentage du montant total du chiffre d'affaires couvert par l'ensemble des unités se chevauchant comparativement au montant total du chiffre d'affaires couvert par l'ensemble des unités sélectionnées par la Méthode I.

Les premiers résultats (voir le tableau 3) montrent comment le modèle SeleMix – qui exploite le comportement longitudinal des entreprises en utilisant le chiffre d'affaires du trimestre T-4 comme covariable – identifie un sous-ensemble de données qui sont signalées par la méthode I. Plus précisément, le pourcentage de données communes identifiées comme étant des erreurs influentes (potentiellement) se situe autour de 32 % du nombre total d'unités sélectionnées par la méthode I. Mais, en moyenne, les unités se chevauchant expliquent les 85 % du nombre de variables cibles contrôlées/vérifiées aux fins de détection d'erreurs. Cela signifie que la méthode de vérification sélective basée sur SeleMix peut probablement être utilisée comme instrument supplémentaire aux fins de détection des erreurs « les plus dangereuses » (pour ce qui est de la variable cible, à savoir le chiffre d'affaires) parmi celles identifiées au moyen de la méthode actuelle.

Tableau 3.
Distribution des données influentes – année 2018

Trimestre	Erreurs influentes			
	Méthode actuelle (I)	SeleMix (II)	I ∩ II	%(I ∩ II sur I)
1	370	195	109	29,5
2	403	228	138	34,2
3	343	190	105	30,6
4	391	209	127	32,5
Total	1 507	822	479	31,8

Ainsi, en résumé, l'hypothèse d'une nouvelle stratégie de V et I pourrait consister à traiter de façon interactive les unités détectées à la fois par les méthodes de vérification sélective I et II, et à traiter automatiquement les 70 % restants d'unités (liés aux 15 % de la variable cible contrôlée). Cette stratégie devrait être plus efficace que la stratégie actuelle, bien qu'elle nécessite trois méthodes qui peuvent être aussi coûteuses, voire plus. Néanmoins, les résultats obtenus sont encourageants quant à la possibilité de réduire le nombre de données à contrôler de façon interactive, tout en réfléchissant à d'autres méthodes automatiques pour les erreurs restantes. La possibilité de vérifier interactivement moins de données tout en maintenant la quantité donnée de la variable cible totale validée soigneusement est la principale piste que nous proposons à partir des premiers résultats.

Dans la section suivante, nous présentons d'autres analyses visant à concevoir une stratégie de V et I plus efficace. En particulier, nous nous sommes intéressés à la façon d'utiliser les méthodes d'apprentissage automatique à la fois pour prédire les unités représentant des données « douteuses » et pour explorer les données actuelles afin de relever des règles et des tendances cachées.

4. Modèles de forêts aléatoires, une proposition pour l'analyse de stratégies de vérification sélective

La méthode d'apprentissage automatique que nous avons utilisée pour la division 46 de la NACE dans la FAS (FAS-NACE 46) repose sur des modèles de forêt aléatoire. Dans un premier temps, pour exécuter un modèle de forêt aléatoire, il faut construire les deux ensembles de données d'entraînement/validation et de données de test. Dans notre étude, nous avons utilisé les 16 000 observations, environ, de 2018 comme ensemble d'entraînement/validation, et 3 000 observations du deuxième trimestre de l'année 2019 comme ensemble de test.

Pour concevoir cette expérience d'utilisation de l'AA dans le contexte de la V et I afin de produire des SCT, nous avons choisi la variable cible et l'ensemble de variables auxiliaires de façon à atteindre les objectifs suivants :

- apprendre un modèle qui relie les entrées (variables de contrôle/covariables) à une variable cible (variable réponse);
- appliquer le même modèle à de nouvelles données afin de prédire de nouveaux cas;
- sélectionner des intrants utiles, habituellement choisis selon la redondance et la non-pertinence;
- extraire les règles de vérification et les tendances cachées d'erreurs potentielles dans les données;
- pour les problèmes de forage de données qui sont souvent caractérisés par une cardinalité importante (à la fois des variables et des unités), le choix des variables d'entrée les plus pertinentes est fait pour ce qui est de la redondance et de la non-pertinence;
- optimiser la complexité : choisir entre des modèles concurrents; la sélection d'un modèle implique toujours un compromis entre le biais (sous-ajustement) et la variance (surajustement).

Aux fins de l'étude, nous construisons une nouvelle variable d'intérêt décrivant si une unité est erronée ou non. En particulier, la variable cible est représentée par le vecteur pour toutes les unités i, observées à plusieurs trimestres T, avec l'indicateur des erreurs influentes de la méthode I actuelle (donnée par l'intersection des procédures A et B, voir le tableau 2) : Yi,T = 1 si l'unité i a eu une influence selon la méthode I; 0 sinon. Les covariables sont représentées par : le chiffre d'affaires aux trimestres T et T-4;
l'emploi aux trimestres T et T-4;
le taux de croissance du chiffre d'affaires de T-4 à T.

On a défini différents modèles de forêt aléatoire (FA) en utilisant différentes variables de base, pour tester différentes hypothèses concernant la façon de prédire les données influentes sur les nouvelles données :

- Modèle 1 : FA avec seulement les variables de base;
- Modèle 2 : FA avec variables de base + indicateur d'erreurs influentes par la Procédure A;
- Modèle 3 : FA avec variables de base + indicateur d'erreurs influentes par la Procédure B;
- Modèle 4 : FA avec variables de base + indicateur d'erreurs influentes par la procédure SeleMix.

Le modèle 1 « apprend » la classification des unités erronées/non erronées à partir des résultats de la méthode I (procédure A et B) seulement sur les données historiques. Cela suppose de ne pas exécuter d'autre procédure de vérification sélective sur les nouvelles données pour prédire la variable de résultat. Les modèles 2, 3 et 4 « apprennent » aussi la classification des unités erronées/non erronées à partir des résultats de l'application de procédures de vérification sélective supplémentaires (Procédure A ou Procédure B ou SeleMix). Par conséquent, si nous comparons les résultats de cette première expérience d'ajout de SeleMix à la méthode I actuelle, chacun des modèles 2, 3 et 4 supposerait d'exécuter une seule méthode de vérification sélective sur de nouvelles données.

5. Résultats et analyse

Au début, les résultats de la matrice de confusion pour chacun de ces modèles – rapportés dans le tableau 5-1 – montrent que pour le modèle 1 de FA, sans l'application d'une méthode de vérification sélective sur les nouvelles données, la phase de test indique une erreur attendue potentielle de 8,1 %. De plus, le modèle 3 de FA produit les meilleurs résultats également sur le test, comme on pouvait l'attendre à partir de la phase d'estimation dans

l'ensemble d'entraînement/validation. Nous pouvons par conséquent avancer qu'une nouvelle phase de V et I devrait inclure uniquement la procédure B pour identifier les erreurs influentes de manière efficace.

Tableau 5-1
Matrice de confusion du modèle sur l'ensemble d'entraînement

	Pourcentage d'erreur	
	Ensemble d'entraînement/validation	Ensemble de test
Modèle 1 FA	6,9	8,1
Modèle 2 FA	5,6	6,7
Modèle 3 FA	1,2	2,0
Modèle 4 FA	6,5	8,1

On effectue d'autres analyses pour mieux déterminer le « type » d'unités prédites comme étant influentes par chaque modèle différent. Le tableau 5-2 indique le nombre d'unités prédites comme erreurs influentes par les modèles de FA 1, 2, 3 et 4 parmi les unités identifiées comme erreurs influentes par la méthode I, dans l'ensemble d'entraînement/ validation. Le modèle 3 FA affiche de bonnes performances : il identifie en effet 1 423 unités sur 1 501. Néanmoins, il impliquerait de traiter de façon interactive un plus grand nombre d'unités (1 526) que ce que nous ferions avec la méthode I actuelle.

Tableau 5-2
Prédiction des erreurs influentes par la méthode I et par les modèles de FA. Nombre d'unités

	Prédiction								
	Modèle 1 FA		Modèle 2 FA		Modèle 3 FA		Modèle 4 FA		Total
Method I	Non influ.	Influ.	Non influ.	Influ.	Non influ.	Influ	Non influ.	Influ.	Influentes
Non influ.	13 977	468	14 023	422	14 342	103	13 996	449	14 445
Influentes	638	863	468	1 033	78	1 423	664	837	1 501
Total	14 615	1 331	14 491	1 455	14 420	1 526	14 660	1 286	15 946

Pour compléter l'analyse, nous étudions la quantité de la variable cible que chaque modèle garantirait de valider. Le tableau 5-3 montre le pourcentage d'unités et de chiffre d'affaires qui serait corrigé dans chaque modèle de FA sur le pourcentage total d'unités et de chiffre d'affaires qui est corrigé par la Méthode I. À titre d'exemple, en utilisant le modèle 1 FA, nous devrions contrôler interactivement 88 % des unités contrôlées par la Méthode I, et nous corrigerions 95 % du montant total du chiffre d'affaires corrigé au moyen des résultats de la méthode I.

Tableau 5-3
Prédiction des erreurs influentes par la méthode I et par les modèles de FA. Pourcentage du nombre d'unités et de la variable cible (chiffre d'affaires).

	Méthode I	Prédiction							
		Modèle 1 FA		Modèle 2 FA		Modèle 3 FA		Modèle 4 FA	
		N^{bre} unités	Chiffre d'affaires	N^{bre} unités	Chiffre d'affaires	N^{bre} unités	Chiffre d'affaires	N^{bre} unités	Chiffre d'affaires
N^{bre} unités	1 501	1 331		1 455		1 526		1 286	
%		88	95	96	101	101	103	89	99

Le tableau 5-3 nous permet de voir que le modèle 1 FA et le modèle 4 FA impliqueraient d'analyser beaucoup moins d'unités, mais avec un pourcentage élevé de contrôle du chiffre d'affaires total. De plus, nous constatons que 825 unités sont identifiées par toutes les méthodes et que 322 unités sont identifiées par le modèle 1 FA et le modèle 4 FA, mais pas par le modèle 3 FA. Par conséquent, l'exécution du modèle 1 FA et du modèle 4 FA, c'est-à-dire l'utilisation d'un modèle de forêt aléatoire avec des variables de base sans aucune procédure de vérification sélective ou en plus des résultats de la procédure SeleMix, est probablement plus efficace pour prédire les unités influentes en nombre d'unités, en permettant de vérifier et de corriger une quantité similaire de la variable cible. Une piste consisterait alors

à déterminer le type d'unités que le modèle 1 FA et le modèle 4 FA identifient pour évaluer si la détermination de leurs caractéristiques peut fournir de l'information utile à la conception d'une procédure de V et I beaucoup plus efficace.

Ces résultats donnent à penser que la méthode actuelle pourrait être remplacée par une autre, qui garantirait l'identification et le traitement des données les plus dangereuses.

6. Conclusions et prochaines étapes

Notre recherche permet de dire que dans le cas des erreurs non dues à l'échantillonnage, la comparaison entre stratégies de V et I n'est pas simple!

Dans l'enquête à court terme FAS sur la division 46 de la NACE, l'utilisation de trois méthodes de vérification sélective pourrait encore représenter une quantité considérable de travail étant données les contraintes de ressources humaines et d'actualité des données. En vue de concevoir une nouvelle stratégie de V et I, nous avons étudié une méthode d'apprentissage automatique utilisant des modèles de forêt aléatoire pour savoir comment les données historiques des processus de V et I peuvent contribuer à l'identification des erreurs influentes.

En résumé, les résultats présentés dans l'article mènent à deux grandes idées pour accroître l'efficacité des stratégies de vérification sélective. Ces deux idées, que voici, devront faire l'objet d'analyses plus approfondies :

a. On pourrait gagner en efficacité au moyen d'une nouvelle procédure basée sur une combinaison de la Procédure B – la plus efficace pour ce qui est de la prédiction des « vraies » erreurs – et de la Méthode II basée sur SeleMix, qui semble se concentrer sur les unités les plus dangereuses et qui donne l'ordre des unités selon le risque qu'elles présentent d'être de « vraies » erreurs influentes;
b. par ailleurs, on pourrait prédire des données douteuses en utilisant uniquement la procédure B, conjuguée à un modèle de forêt aléatoire approprié.

À cette étape, il faudrait poursuivre les recherches afin de mieux définir le mode d'utilisation de l'apprentissage automatique dans un but différent, à savoir pour comprendre pourquoi les différents modèles de FA se comportent différemment, et ainsi découvrir si l'information disponible comporte des tendances cachées permettant d'identifier les données influentes. Une analyse plus approfondie et d'autres études expérimentales s'appuyant sur un plus grand nombre de données historiques sont prévues. Elles devraient comparer différentes méthodes de vérification sélective dans l'enquête FAS et évaluer différents procédés de V et I.

Ces travaux devraient clarifier les idées sur les changements dans les modèles de vérification sélective et, de façon plus générale, dans le processus de V et I, susceptibles de considérablement améliorer des aspects opérationnels de l'ensemble du processus de production statistique.

Bibliographie

Beck M., F. Dumpert et J. Feuerhake (2018), Machine Learning in Official Statistics.

Breiman L. et A. Cutler (2001). Random Forests for Classification and Regression, disponible à l'adresse : https://cran.r-project.org/web/packages/randomForest/index.html.

Buglielli, M.T., and U. Guarnera (2016), SeleMix: Selective Editing via Mixture Models, Version 1.0.1, disponible à l'adresse : https://CRAN.R-project.org/package=SeleMix.

Di Zio M., and U. Guarnera (2013), Contamination Model for Selective Editing. *Journal of Official Statistics*, Vol. 26, n. 4, p. 539-556.

Guarnera U., and M.T. Buglielli (2013), SeleMix: an R Package for Selective Editing, disponible à l'adresse : https://www.istat.it/it/files/2014/03/SeleMix-vignette.pdf.

Luzi O., T. De Waal, B. Hulliger, M. Di Zio, J. Pannekoek, D. Kilchmann, U. Guarnera, J. Hoogland, A. Manzari and C. Tempelman (2008), *Recommended practices for editing and imputation in cross-sectional business surveys*. Eurostat.

CEE-ONU (2015), Generic Statistical Data Editing Models - GSDEMs, Version 1.0, octobre 2015, disponible à l'adresse : https://statswiki.unece.org/display/sde/GSDEMs.