Direction générale
Direction de la méthodologie et de la coordination statistique et internationale
Département « méthodes statistiques »
Division « recueil et traitement de l'information »

Dossier Suivi par :
LEROY Theo
Tél : 0187695533
Mèl : theo.leroy@insee.fr

HIMPENS Stéphanie
Tél : 0187695516
Mèl : stephanie.himpens@insee.fr

MALHERBE Lucas
Tél : 0187695578
Mèl : lucas.malherbe@insee.fr

Note à l'attention des destinataires in fine

Montrouge, le 8 février 2021
N°2021_3256_DG75-L401

Objet : Tirage d'échantillons pour l'opération d'annotation du RP en PCS 2020

L'adoption de la nouvelle nomenclature PCS 2020 pose la question de la codification automatique des bulletins individuels de recensement de la population. Pour la PCS 2003, cette tâche était assurée par l'outil Sicore et suivie d'une phase de reprise manuelle. Suite à la rénovation de la PCS, l'absence d'une méthode opérationnelle pour coder invite à concevoir des modèles de classification supervisée qui sembleraient plus appropriés qu'une base de connaissance Sicore pour le traitement chaque année d'environ trois millions de bulletins individuels. Ce type de modèle apprend à classer les observations grâce à un ensemble d'exemples annotés. Il faut donc fournir au modèle une « base d'apprentissage » contenant des exemples de bulletins individuels ainsi que la PCS qui leur correspond pour que celui-ci soit en mesure de fournir une prédiction sur de nouvelles observations. Afin de constituer cette base d'apprentissage nécessaire pour l'initialisation du modèle, une campagne d'annotation est en cours de préparation. Elle aura lieu au premier semestre 2021.

La note n°2021_2078_DG75-F520 fixe à 120 000 le nombre de bulletins individuels qui pourront être annotés manuellement (incluant pour chacun un double codage ainsi qu'un arbitrage en cas de divergence des deux premiers codages). Ces 120 000 bulletins individuels doivent donc être sélectionnés au préalable parmi l'ensemble des données disponibles.

L'approche la plus simple consisterait à tirer les 120 000 bulletins individuels via un sondage aléatoire simple à probabilités d'inclusion égales. L'objectif de cette note est d'examiner plusieurs solutions alternatives de tirage d'échantillon. Le but de cette réflexion est double. Il s'agit d'une part de maximiser le gain d'information compte tenu de la contrainte sur le nombre d'annotations possible tout en permettant d'autre part d'évaluer au mieux les performances du modèle.

Trois types de professions sont collectées dans le recensement : profession actuelle des salariés (PROFS), profession actuelle des non-salariés (PROFI) et profession antérieure (PROFA). Le questionnaire et donc les variables collectées sont propres à chaque type de profession. Il est ainsi envisagé d'entraîner trois modèles différents. Cette note propose une ventilation du volume de données à annoter entre ces trois types de professions dans le but d'homogénéiser la précision entre les trois futurs modèles.

La cheffe de l'unité « SSP Lab »	La cheffe de la division « recueil et traitement de l'information »
Signé : Elise Coudin	Signé : Amandine Schreiber

Liste de diffusion

Transmission à :

DSDS/DMTR : Souheil BENMEBKOUT, Maxime EXAVIER, Gwennaël SOLARD
DR69/SeRN : Caroline ANGUIER, Pascal ARDILLY, Mireille BEL, Philippe BERTRAND, Carlos PORTAS
DSDS/Division Emploi : Olivier CHARDON, Élodie PEREIRA, Chloé TAVAN
DR25/Pôle Expertise et reprise PCS : Éric HANRIOT, Laurence LABOSSE, Florent MAIRE, Audrey MIRAULT
SSP Lab

Copie à :

DMCSI : Sylvie LAGARDE, Patrick SILLARD,
DSDS : Valérie ROUX

1. Stratégies de sélection d'un échantillon d'entraînement

Des tests préliminaires ont montré qu'il était plus intéressant de faire apprendre le modèle sur des bulletins individuels tous différents pour obtenir une base d'apprentissage la plus diversifiée possible. Ainsi, dans tous les scénarios, on procède d'abord à une suppression des doublons (libellé x variables annexes) afin d'éviter de faire annoter deux fois le même bulletin.

Quatre scénarios ont été étudiés. Pour chacun des scénarios et pour chaque type de profession, des éléments de performance sont indiqués. Ceux-ci ont été calculés à partir des labels en nomenclature PCS 2003 obtenus dans les enquêtes annuelles de recensement (EAR) passées c'est-à-dire via Sicore ou le processus de reprise manuelle appelé Recap.

Scénario 1 : Tirage aléatoire avec probabilités d'inclusion proportionnelles à la fréquence d'apparition des caractéristiques individuelles

Ce scénario consiste à privilégier dans l'échantillon d'apprentissage les observations que l'on retrouve le plus fréquemment dans les bulletins individuels. Il s'agit d'un tirage systématique avec probabilités d'inclusion inégales (proportionnelles à la fréquence d'apparition des mêmes caractéristiques individuelles dans l'EAR). Au préalable, il faut identifier l'ensemble des variables qui seront exploitées pour regrouper les bulletins individuels. Ce sont celles qui ont été identifiées comme étant les plus porteuses d'information pour coder dans la nomenclature des professions. Les variables retenues sont listées dans le tableau ci-après.

Professions	**Variables de tirage**
Profession actuelle des salariés (PROFA)	• Libellé déclaré de la profession actuelle normalisé (suppression des caractères de ponctuation et des mots vide de sens) • Statut professionnel déclaré • Position professionnelle déclarée • Secteur d'activité de l'établissement employeur en NAF5
Profession actuelle des non-salariés (PROFI)	• Libellé déclaré de la profession actuelle normalisé (suppression des caractères de ponctuation et des mots vide de sens) • Statut professionnel déclaré • Pluralité déclarée de salariés employés • Tranche d'effectif de l'entreprise (appariement Sirene) • Secteur d'activité de l'établissement employeur en NAF5
Profession antérieure (PROFS)	• Libellé déclaré de la profession antérieure normalisé (suppression des caractères de ponctuation et des mots vide de sens) • Statut professionnel antérieur déclaré

Tableau 1: Ensemble des caractéristiques retenues pour chaque type de profession

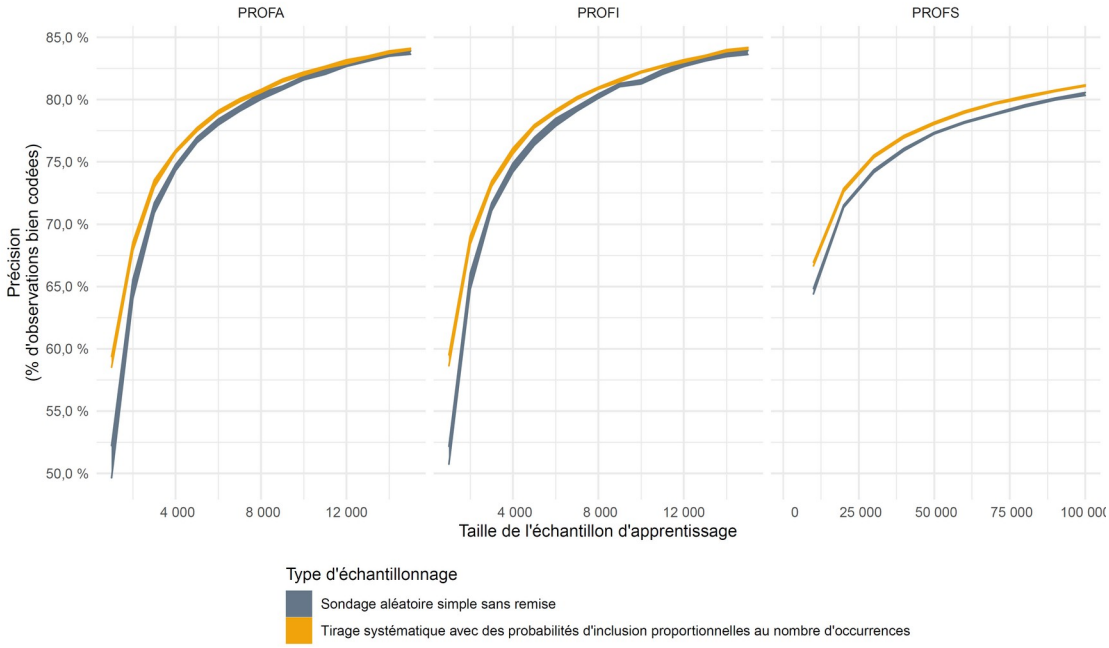


Figure 1: Comparaison de l'efficacité de l'échantillonnage proportionnel à la taille (Scénario 1) et le sondage aléatoire simple avec probabilités d'inclusion égales

Quel que soit le type de profession, ce tirage systématique améliore la précision des modèles par rapport au sondage aléatoire simple où le nombre d'occurrences n'est pas pris en compte.

Scénario 2 : Tirage aléatoire à partir d'un *embedding*

Le scénario 1 s'appuie sur les variables brutes. Les bulletins individuels qui présentent les mêmes valeurs sur le croisement de variables décrit précédemment sont regroupés. Cependant, exploiter ces variables peut conduire à générer des regroupements trop stricts. Deux observations avec des libellés légèrement différents (une faute d'orthographe, une permutation de mots…) sont considérés comme distincts. Il est possible d'utiliser des méthodes de *word-embedding* pour présenter une information textuelle dans un espace vectoriel de dimension finie. Deux mots ou textes sémantiquement proches seront représentés par deux vecteurs numériquement proches. Des méthodes de *clustering* (*K-means*, CAH, DBSCAN…) peuvent être utilisées dans l'espace de *word-embedding* pour constituer n groupes distincts et tirer un représentant selon une loi uniforme dans chaque groupe. Les individus au sein d'un groupe doivent être le plus homogènes possible.

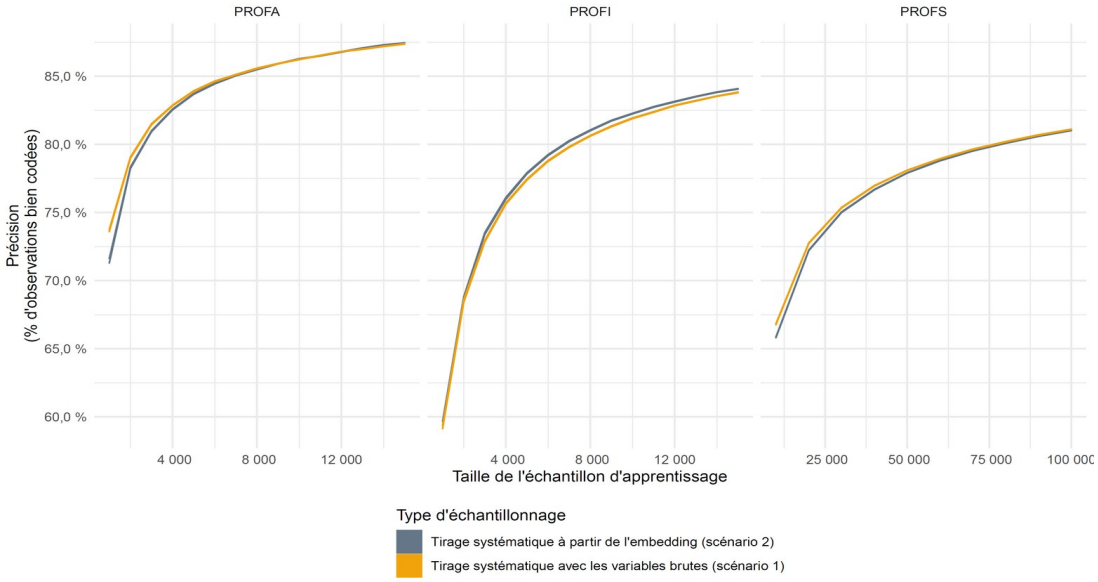


Figure 2: Comparaison de stratégies de tirage proportionnel à la taille à partir des variables d'enquêtes ou d'un embedding (fastText de dimension 100)

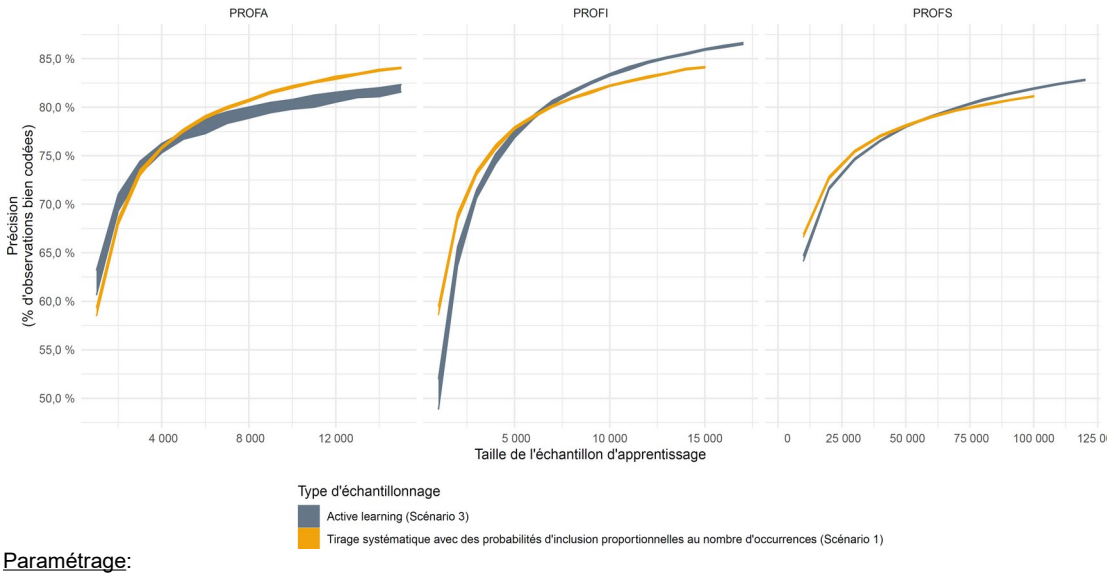
L'utilisation de *l'embedding* améliore la pertinence des unités échantillonnées pour construire un modèle le plus précis possible uniquement sur la profession actuelle des indépendants. De plus, ce type de stratégie soulève quelques incertitudes sur la validité du tirage, car elle repose sur plusieurs partis-pris en amont comme *l'embedding* utilisé (fastText, word2vec, camenbert…), ses paramètres et des hypothèses provenant de la méthode de regroupement choisie.

Au total, le gain de performance de cette méthode par rapport au scenario 1, avec pour autant des paramètres optimisés, apparaît faible vu le nombre d'hypothèses et de choix arbitraires nécessaire à cette méthode, des risques de biais sur l'échantillon de test et la difficulté de combiner un tirage sur *embedding* avec une procédure d'*active learning*. Le temps a manqué pour analyser cette stratégie sur la profession actuelle des salariés et la profession antérieure.

Scénario 3 : Active learning – Tirage d'un petit échantillon puis sélection de nouvelles observations à annoter sur la base d'un indice de confiance

L'*active learning* est une méthode de priorisation des tâches de labellisation au cours de la phase d'annotation. Elle repose sur une interaction entre l'algorithme d'apprentissage et l'état courant du processus de labellisation. Au fur et à mesure de l'annotation, c'est le modèle lui-même qui va servir à prioriser les bulletins individuels à faire annoter afin de maximiser le gain marginal d'information et donc les performances du modèle.

En pratique, un premier échantillon de petite taille $n_{initial}$ est tiré (par sondage aléatoire simple, tirage systématique sur les variables initiales ou d'*embedding*). Une fois celui-ci annoté, un premier modèle est entraîné. Puis, une prédiction à partir de ce modèle est réalisée sur l'ensemble des données de la base de sondage excepté l'échantillon. Enfin, les n_{pas} observations dont les prédictions sont les plus incertaines sont sélectionnées. Pour mesurer le degré d'incertitude, il est d'usage de considérer la différence de probabilité entre les deux classes estimées les plus probables par le modèle. Ce processus de sélection des n_{pas} observations les plus pertinentes à annoter à chaque étape puis d'annotation est répété jusqu'à atteindre la taille n d'échantillon fixée au préalable.



Paramétrage:

	Méthode de tirage de l'échantillon initial	Taille de l'échantillon initial $n_{initial}$	Pas de l'*active learning* n_{pas}
PROFA	Sondage aléatoire simple	1 000	1 000
PROFI	Sondage aléatoire simple	1 000	1 000
PROFS	Sondage aléatoire simple	10 000	10 000

Figure 3: Comparaison des stratégies d'active learning (priorisation au cours de l'annotation) et de tirage systématique (priorisation en amont de l'annotation)

Au travers des expérimentations menées, il apparaît que la stratégie d'*active learning* est meilleure pour des volumes de données plus grands. Le volume de données à partir duquel ce scénario est plus performant que le premier peut être diminué en abaissant très fortement le pas de l'active learning (par exemple toutes les 5 ou 10 observations). Cependant, nous ne serions pas en mesure de mettre en œuvre ce scenario avec un pas trop petit car cela demanderait au sein de l'application de labellisation de recalculer les modèles très régulièrement. L'intégration de cette fonctionnalité soulèverait des questions délicates d'implémentation qui ne pourraient être résolues dans un délai court. Par exemple, l'entraînement avec les quelques observations supplémentaires pourrait être plus longue que

l'annotation manuelle de ces observations, comment éviter alors des phénomènes de latence dans l'annotation ?

De plus, l'*active learning* suppose la définition, en amont, d'une méthode d'apprentissage supervisée (machine à vecteur de support, réseau de neurones artificiels, arbre de décision etc.) afin de prioriser les documents à annoter. Bien que fastText soit un classifieur très performant au vu des travaux menés sur l'ancienne nomenclature, privilégier certains documents à annoter selon cet *a priori* pourrait sembler abusif. De plus, l'analyse ex-post du meilleur modèle sera très certainement faussée et favorisera à l'excès fastText car, en amont, les observations les plus pertinentes pour entraîner ce dernier auront été sélectionnées. En revanche, l'approche par *active learning* pourrait être tout à fait adaptée en production pour les campagnes suivantes (après première estimation du modèle).

Scénario 4 : Échantillon principal et réserve déterminée par *active learning*

Étant donné les difficultés d'intégration du scénario précédent et des *a priori* sur lesquels il repose, il est possible de l'adapter pour tenter de profiter des avantages attendus sur la précision des modèles.

Un échantillon principal d'une taille $n_{principal}$ est au préalable sélectionné par tirage systématique (scénario 1 ou 2). Une fois ce dernier annoté, un modèle est entraîné à partir des données de l'échantillon principal et on sélectionne $n_{reserve}$ nouvelles observations dans la base de sondage à partir d'un indice de confiance dans la codification. Il s'agit du scénario 3 dans lequel une seule itération d'*active learning* a été réalisée. La mise en place d'une telle procédure est moins lourde que pour le scénario précédent et de surcroît elle n'oblige pas à sélectionner une classe de modèles de façon trop précoce.

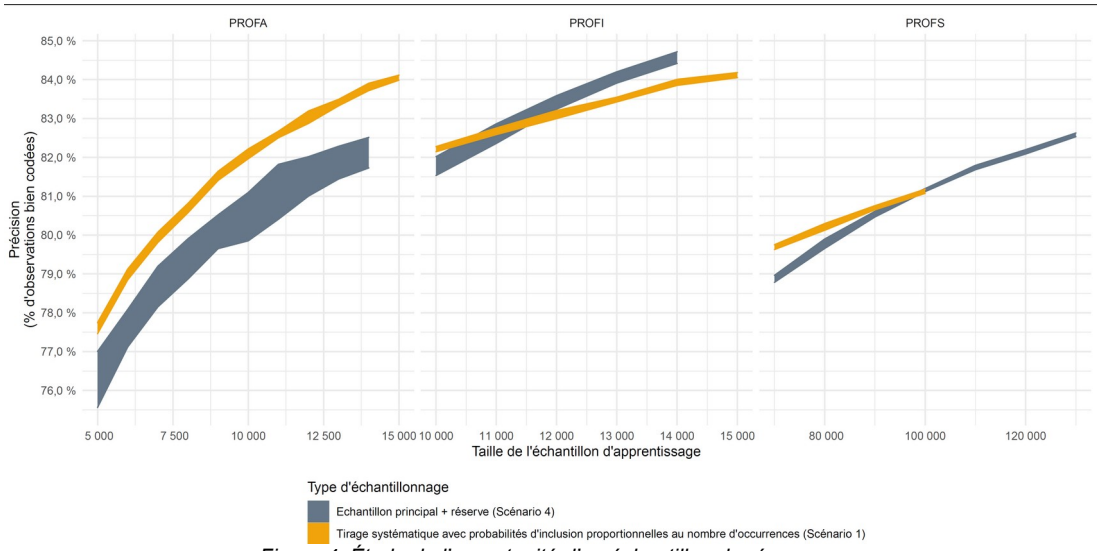


Figure 4: Étude de l'opportunité d'un échantillon de réserve

Cette stratégie gagne en performance avec le volume de données, elle ne dépasse le scenario 1 que lorsqu'un volume important de données est atteint Pour les premiers bulletins à annoter, la priorisation par la fréquence d'apparition est plus pertinente que la sélection grâce à l'indice de confiance du modèle car les prédictions de celui-ci sont encore instables.

2. Stratégies d'évaluation des modèles

Il faut être particulièrement prudent en apprentissage supervisé sur l'évaluation de la qualité des prédictions du modèle. En effet, utiliser les mêmes données pour l'apprentissage et l'évaluation conduit en général à une surestimation des performances réelles du modèle. L'approche classique consiste donc à séparer le jeu de données en deux en créant un échantillon d'entraînement et un échantillon de test, ce dernier permettant d'évaluer les performances en « conditions réelles », c'est-à-dire sur des données que le modèle n'a jamais rencontrées durant la phase d'apprentissage.

Dans le cadre de la PCS 2020, deux types de stratégies sont envisageables.

Scénario A : Échantillons d'entraînement et de test séparés

L'échantillon de test S_{test} peut être déterminé par sondage aléatoire simple à probabilités égales dans l'ensemble des bulletins individuels pour constituer un ensemble représentatif. Dans ce cas, la proportion d'observations bien classées dans l'échantillon permet d'évaluer la performance du modèle. Il est également possible de constituer un échantillon par tirage systématique où les probabilités d'inclusion sont proportionnelles au nombre d'occurrences. Cette méthode est identique à celle décrite dans le scénario 1 pour l'échantillon d'apprentissage. Dans ce cas, pour évaluer la performance du modèle, la proportion d'observations bien classées τ doit être estimée en pondérant par l'inverse des probabilités d'inclusion; ce qui conduit à l'estimateur approximativement sans biais ci-dessous :

$$\hat{\tau}_{test} = \frac{\sum_{i \in S_{test}} \frac{t_i 1\left(y_i = \hat{y}_i^{(modele)}\right)}{\pi_i}}{\sum_{i \in S_{test}} \frac{t_i}{\pi_i}}$$

où

- $1(y_i = \hat{y}_i^{(modele)})$ est l'indicatrice de bonne prédiction (vaut 1 si le modèle entraîné sur les données d'apprentissage prédit bien la valeur attendue et 0 sinon) ;
- t_i est le nombre d'occurrences de l'observation i dans la base de sondage (EAR 2020). C'est le nombre de répétitions du même n-uplet (libellé x variables annexes) ;
- π_i est la probabilité d'inclusion de l'observation i dans l'échantillon de test.

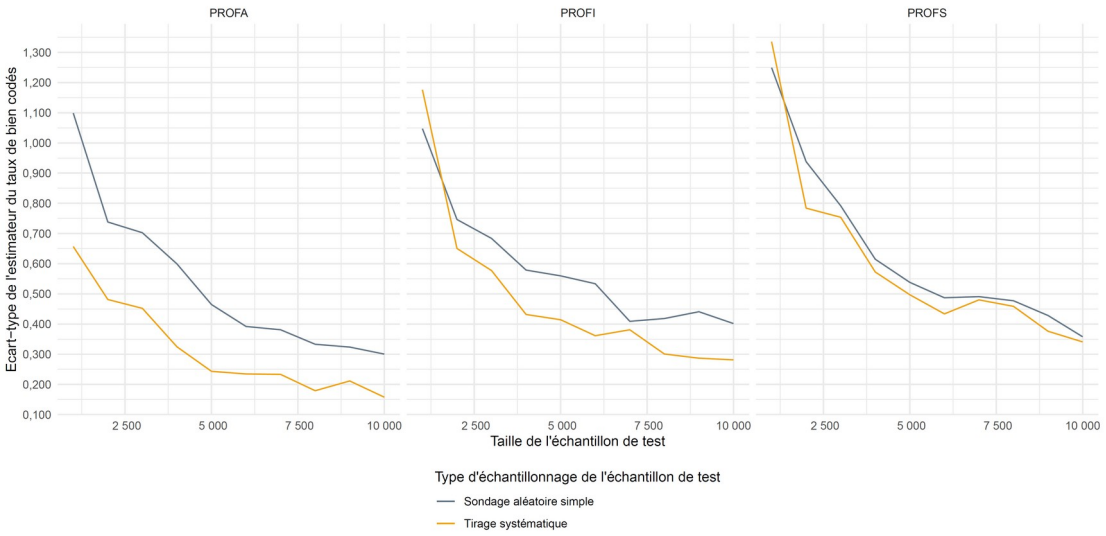


Figure 5: Comparaison de scénarios d'échantillonnage pour le test (entraînement des modèles à partir du scénario combinant un échantillon principal et une réserve)

Comme le montre la figure 5, le tirage systématique est plus performant que le sondage aléatoire simple, à taille d'échantillon identique, car il permet de diminuer la variance de l'estimateur du taux d'observations bien codées par les modèles.

Une fois l'échantillon de test tiré, l'échantillon d'entraînement est tiré selon l'un des quatre scénarios proposés dans la partie précédente.

Scénario B : Échantillon commun avec validation croisée

Il est possible d'exploiter l'échantillon d'entraînement dans le but d'évaluer le modèle, on procède alors par validation croisée. L'échantillon d'apprentissage est découpé équitablement en K groupes appelés « folds ». Les données peuvent par exemple être réparties en 100 folds. Pour chaque fold k, un modèle est entraîné à partir des données des $K-1$ autres folds puis la précision est évaluée en mobilisant les observations du fold k. La précision τ sur le fold k est estimée de la façon suivante :

$$\hat{\tau}_k = \frac{\sum_{i \in fold_k} \frac{t_i 1(y_i = \hat{y}_i^{(modele_{\overline{fold_k}})})}{\pi_i}}{\sum_{i \in fold_k} \frac{t_i}{\pi_i}}$$

où

- $1(y_i = \hat{y}_i^{(modele_{\overline{fold_k}})})$ est l'indicatrice de bonne prédiction (vaut 1 si le modèle entraîné sur les données n'appartenant pas au fold k a bien prédit la valeur attendue et 0 sinon) ;
- t_i est le nombre d'occurrences de l'observation i dans la base de sondage (EAR 2020). C'est le nombre de répétitions du même n-uplet (libellé x variables annexes) ;
- π_i est la probabilité d'inclusion de l'observation i dans l'échantillon pour l'apprentissage.

La précision globale peut alors être estimée par la moyenne des précisions sur les K folds. :

$$\hat{\tau}_{CV} = \frac{\sum_{k=1}^{K} \hat{\tau}_k}{K}$$

Le scénario B est moins coûteux car les données pour entraîner les modèles et les comparer sont mutualisées. Toutefois, plusieurs inconvénients doivent être mentionnés :

- La comparaison des modèles candidats est plus longue car elle nécessite de répéter le processus d'entraînement puis de prédiction K fois.
- Cette stratégie n'est pas compatible avec un scénario d'échantillonnage pour lequel les probabilités d'inclusion ne sont pas connues. C'est le cas des scénarios 2 et 3.
- L'estimateur par validation croisée risque d'être biaisé pour des petites tailles d'échantillon. En effet, si l'échantillon est petit, les observations seront très différentes les unes des autres et auront donc toutes une contribution forte dans le modèle appris. En retirer une partie pour évaluer et apprendre sur le reste conduira à une sous-estimation de la précision.

3. Synthèse et ventilation des données entre les trois types de professions

Déterminer la répartition optimale entre les trois types de professions est délicat car il n'est pas possible de tester toutes les combinaisons de volumétries sur tous les scénarios d'échantillonnage. À défaut de quadriller l'ensemble des possibles, des propositions ont été testées successivement en tâtonnant dans le but de maximiser le taux d'observations bien codées pour les trois types de professions. L'objectif fixé par le département de la démographie est d'aboutir à une fiabilité (taux de bien codés) supérieure à 80 % pour les bulletins individuels codés automatiquement (sur liste ou via un modèle d'apprentissage statistique). L'efficacité (taux de codage automatique) doit être supérieure à 87 %.

Par ailleurs, même si nous avions eu la possibilité de tester toutes les combinaisons, il aurait été excessif de qualifier la solution la plus prometteuse d'optimale car l'expérimentation a été menée dans un contexte différent : les codes sont issus du processus actuel (Sicore+reprise manuelle) et la nomenclature utilisée était la PCS 2003. L'ancienne nomenclature comportait plus de catégories (486 contre 311 dans la PCS 2020) donc la tâche de classification était plus complexe. Le travail en amont nous permet plutôt d'identifier des stratégies d'échantillonnage (méthode d'échantillonnage et taille pour les données d'entraînement et de test) viables et suffisamment efficaces car elles permettraient d'améliorer la pertinence des documents à annoter par rapport à un sondage aléatoire simple.

De plus, dans le cadre d'une évaluation des performances sur un échantillon de test, la répartition entre données d'apprentissage et de test est complexe. Plus les données seront mobilisées dans l'échantillon d'apprentissage et plus les modèles seront précis. Toutefois, moins il y aura de données disponibles pour tester les modèles, plus les intervalles de confiance pour l'estimation du taux de bulletins individuels bien codés seront grands.

Profession actuelle des salariés (PROFS)

		Intervalle de confiance minimal*	Stratégie pour l'évaluation	
			Scénario A (5 000 BI de test)	Scénario B (K=100)
	Scénario 1 $n=90\,000$	79,9 ± 0,2	79,9 ± 1,0	80,5 ± 0,2

Stratégie pour l'apprentissage	Scénario 3 $n=90\,000$ $n_{initial}=5000$ $n_{pas}=5\,000$	80,0 ± 0,2	80,2 ± 0,9	
	Scénario 4 $n=90\,000$ $n_{principal}=70\,000$ $n_{reserve}=20\,000$	80,3 ± 0,1	80,2 ± 0,8	80,6 ± 0,4

Profession actuelle des non-salariés (PROFI)

		Intervalle de confiance minimal*	Stratégie pour l'évaluation	
			Scénario A (2 000 BI de test)	Scénario B (K=100)
Stratégie pour l'apprentissage	Scénario 1 $n=12\,500$	84,0 ± 0,3	83,8 ± 1,4	82,6 ± 0,4
	Scénario 3 $n=12\,500$ $n_{initial}=500$ $n_{pas}=500$	84,7 ± 0,7	85,4 ± 1,2	
	Scénario 4 $n=12\,500$ $n_{principal}=10\,000$ $n_{reserve}=2500$	83,7 ± 0,4	84,2 ± 1,5	82,1 ± 0,4

Profession antérieure (PROFA)

		Intervalle de confiance minimal*	Stratégie pour l'évaluation	
			Scénario A (2 000 BI de test)	Scénario B (K=100)
Stratégie pour l'apprentissage	Scénario 1 $n=7\,500$	88,5 ± 0,1	88,5 ± 0,9	82,2 ± 1,1
	Scénario 3 $n=7\,500$ $n_{initial}=500$ $n_{pas}=500$	81,4 ± 1,0	81,5 ± 1,3	
	Scénario 4 $n=7\,500$ $n_{principal}=6500$ $n_{reserve}=1000$	88,4 ± 0,3	88,4 ± 1,0	82,1 ± 0,5

Tableau 2 : Estimations d'intervalles de confiance à 95 % du pourcentage de bulletins individuels bien codés selon différentes stratégies

* : Intervalle de confiance estimé sur un jeu de données de test très grand (20 % des données de l'EAR 2019 et disjoint des données d'apprentissage) par Monte Carlo. Par la loi des grands nombre, cet intervalle de confiance est centré sur l'espérance de la précision des modèles entraînés sur cette taille et ce type d'échantillon. On peut le qualifier de minimal car étant donnée la taille des données exploitées pour l'évaluation, on peut supposer que seule la variabilité de l'échantillon d'apprentissage entre en compte dans le calcul de cet intervalle de confiance.

Compte tenu des éléments théoriques, de considérations pratiques sur la mise en place des différents scénarios ainsi que des résultats des expérimentations décrites dans ce document, deux scénarios nous paraissent se détacher pour le tirage de l'échantillon d'entraînement :

- Le scénario 1 (tirage systématique avec probabilités d'inclusion proportionnelle à la fréquence d'apparition) est le plus facile à mettre en œuvre car il ne nécessite qu'une seule étape de tirage en amont de la campagne d'annotation. Il permet d'atteindre des performances comparables aux autres scénarios, et en tout cas meilleures qu'avec un tirage naïf par sondage aléatoire simple.
- Le scénario 4 (tirage systématique + réserve constituée par *active learning*) consiste en deux étapes distinctes mais il est susceptible dans certains cas d'améliorer la qualité de l'échantillon bâti. De plus, dans le cas où il serait impossible d'annoter 120 000 bulletins individuels, notamment du fait des moyens humains disponibles, il est intéressant de diviser l'échantillon d'apprentissage en un bloc principal et une réserve.

Nous retenons donc le scénario 4 pour le tirage de l'échantillon d'apprentissage.

En ce qui concerne la stratégie d'évaluation, l'utilisation d'un échantillon de test disjoint de l'échantillon d'apprentissage (scénario A) nous semble plus adéquate que par validation croisée. En effet, cette dernière méthode risquerait de biaiser les estimations, notamment sur la profession antérieure pour laquelle les volumes sont plus faibles.

La ventilation des 120 000 observations ci-après nous semble ainsi adaptée pour répondre aux attentes du département de la démographie :

	Échantillons		
	Apprentissage		Test
	Bloc principal	Réserve	
Profession actuelle des salariés (PROFS)	70 000	20 000	5 000
Profession actuelle des non-salariés (PROFI)	10 000	2 500	2 000
Profession antérieure (PROFA)	6 000	1 500	2 000

Tableau 3 : Suggestion de ventilation pour la campagne d'annotation dédiée à l'initialisation de modèle supervisés pour le codage des professions "hors-liste"

De plus, mille observations, omises dans le tableau précédent, constitueront un « Gold-standard ». Les données seront issues du pilote de l'enquête Emploi. Les professions auront été codées au préalable en PCS 2020 par le pôle d'expertise PCS. Au début de la campagne d'annotation, ces mille observations seront à nouveau codées en PCS 2020 par les gestionnaires dans les établissements régionaux afin de s'assurer que les consignes de codage ont été comprises.