Journées de méthodologie statistique de l'Insee 2022

JMS | Application de techniques de machine learning pour coder les professions en PCS 2020

Théo Leroy (*), Lucas Malherbe (*), Tom Seimandi (*)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

Mots-clés. PCS, fastText, classification hiérarchique

Domaines. Codification automatique, NLP, réseaux de neurones

Résumé

La nomenclature des professions et catégories socioprofessionnelles (PCS) a subi une rénovation en 2020. Celle-ci s'accompagne de la promotion d'un outil d'autocomplétion des libellés de profession dans une liste de libellés enrichis permettant le codage direct dans un poste de la nomenclature. Jusqu'ici, la codification automatique des professions issues de l'enquête annuelle de recensement était assurée par un moteur de règles déterministes appelé Sicore. Le passage par la liste de libellés enrichis rend caduque cet environnement dès lors que l'outil d'autocomplétion est utilisé. Il n'est pas prévu de développer un moteur de règles complet similaire à Sicore pour les libellés en dehors de la liste de libellés enrichis en PCS 2020. Or, l'outil d'autocomplétion ne sera disponible que pour des collectes informatisées et par ailleurs il comprend la possibilité de répondre « hors liste ».

Pour être en mesure de coder les bulletins papier comme les réponses informatisées « hors liste », un algorithme de codification automatique en PCS 2020 de ces bulletins doit être créé. L'objectif est de maintenir le taux de codifications correctes tout comme le taux d'envoi en reprise manuelle à des niveaux similaires à l'existant.

Il s'agit de proposer, ici, plusieurs solutions à partir de modèles de classification supervisée. Des premiers travaux d'utilisation de techniques de *machine learning* sur la PCS 2003 se sont révélés suffisamment prometteurs pour poursuivre dans cette direction. Suite au report de l'enquête de recensement de 2021, une base de données a été annotée dans la nouvelle nomenclature PCS 2020 lors d'une campagne de labellisation. Cette vaste campagne d'annotation d'environ 120 000 bulletins s'est tenue en double codage avec arbitrage et offre une opportunité unique d'entraîner directement des modèles de *machine learning* à prédire dans cette nouvelle nomenclature, sans recourir à l'implémentation de règles déterministes comme le faisait Sicore.

Les meilleurs modèles identifiés lors des tests menés sur la PCS 2003 ont été répliqués sur ces données pour coder en PCS 2020. Ils font usage du classifieur *fastText*, un réseau de neurones à une couche cachée. Cependant, un travail d'adaptation et de réestimation des modèles est

nécessaire pour plusieurs raisons. D'abord, les nomenclatures PCS 2003 et PCS 2020 présentent des différences non négligeables. Ensuite, les variables mobilisables pour prédire un poste de la nomenclature ne sont pas exactement les mêmes pour les deux nomenclatures.

Les premiers résultats, évalués sur un échantillon de test annoté manuellement et indépendant de l'échantillon d'apprentissage, suggèrent une précision de codage de 67 % pour des libellés de profession actuelle salariée hors champ de l'index des professions utilisé par l'outil d'autocomplétion. Dans le but d'améliorer les performances des modèles, plusieurs axes sont explorés. Un premier vecteur d'amélioration est la prise en compte de la hiérarchie dans la nomenclature en agrégeant des modèles sur plusieurs niveaux de la nomenclature (combinaison linaire des probabilités obtenues à chaque niveau, arbre de modèles conditionnels). Une autre piste est d'augmenter la taille de l'échantillon d'apprentissage avec des informations provenant d'autres enquêtes. Cet article présentera le résultat de ces approches.

Abstract

In this work, we use machine learning models relying on text embeddings to code occupation for observations from France's population census. Observations have both a free text feature describing the individual's occupation and categorical features encoding information on the type of occupation, on the individual's employer, etc. We find that when combined with a reasonably-sized set of rules designed to classify recurring observations, embeddings which take n-grams into account lead to good accuracy results, better than the ones obtained today with a codification engine which only leverages a complex set of decision rules.

Introduction

La nomenclature des professions et catégories socioprofessionnelles (PCS) a subi une rénovation en 2020 [1]. La nouvelle nomenclature part de six groupes très larges, subdivisés en 30 catégories socioprofessionnelles qui sont divisées en 126 professions regroupées, elles mêmes divisées en 316 professions. Avec cette rénovation se pose la question de l'évolution de la codification automatique des bulletins individuels du recensement de la population. Sur chaque bulletin individuel figurent des informations sur la catégorie socioprofessionnelle de l'individu enquêté. Par exemple, dans le cas où ce dernier est salarié au moment où il répond à l'enquête, il doit renseigner un libellé de profession pour répondre à la question « *Quelle est votre profession principale ?* ». Il doit aussi choisir sa *position professionnelle* parmi 9 catégories, dont « *Manoeuvre, ouvrier spécialisé* », « *Ingénieur, cadre d'entreprise* » et « *Agent de catégorie A de la fonction publique* ».

Jusqu'à présent, un moteur de règles déterministes appelé Sicore était utilisé pour déterminer la catégorie socioprofessionnelle dans la nomenclature PCS 2003 à partir de ces informations. Dans certains cas, Sicore ne parvenait pas à choisir une catégorie pour un bulletin individuel. Les bulletins non codés par Sicore étaient ainsi annotés au cours d'une phase de reprise manuelle.

La mise à jour de Sicore en nomenclature PCS 2020 représenterait un travail conséquent. En outre, cet outil de codification est difficile à maintenir sur la durée tant les règles déterministes qui constituent le moteur de traitement sont nombreuses et parfois complexes. Ainsi, l'INSEE a lancé une expérimentation autour de l'utilisation de techniques de *machine learning* pour coder la catégorie socioprofessionnelle de bulletins individuels du recensement. Pour entraîner un modèle de classification supervisée sur cette tâche de codification, il faut lui fournir une « base d'apprentissage » contenant des exemples de bulletins individuels ainsi que la PCS qui leur correspond. Après entraînement, le modèle sera en mesure de fournir une prédiction sur de nouvelles

observations.

Suite au report de l'enquête annuelle de recensement de 2021, une base de données a pu être annotée dans la nouvelle nomenclature PCS 2020 lors d'une campagne de labellisation. Cette vaste campagne d'annotation d'environ 120 000 bulletins s'est tenue en double codage avec arbitrage et offre une opportunité unique d'entraîner directement des modèles de *machine learning* à prédire dans la nouvelle nomenclature.

1 Codification automatique et données

Trois types de professions sont collectées dans le recensement de la population, en fonction de la situation de l'individu enquêté (Figures 11 et 12 en annexe) :

— Dans le cas où l'enquêté est salarié, il renseigne des informations relative à une profession actuelle salariée (PROFS dans la suite de ce document) ;
— Dans le cas où l'enquêté travaille à son compte ou en tant que chef d'entreprise, il renseigne des informations relatives à une profession actuelle non-salariée (PROFI dans la suite) ;
— Dans le cas où l'enquêté ne travaille pas au moment de l'enquête, il renseigne des informations relatives à une profession antérieure (PROFA dans la suite).

Le questionnaire et donc les variables collectées sont en partie propres à chaque type de profession. Par exemple, concernant sa profession antérieure, un enquêté ne doit pas donner d'information sur l'entreprise qui l'employait, contrairement au cas de la profession actuelle pour un enquêté travaillant au moment de l'enquête. De même, un enquêté salarié doit renseigner sa position professionnelle, ce qu'il ne doit pas faire s'il est indépendant. Ainsi, trois tâches de codification différentes (une pour chaque type de profession) sont à assurer pour le recensement de la population. Les variables pertinentes pour chaque tâche de codification sont détaillés en 1.1.

1.1 Données

La variable principale utilisée pour déterminer le code PCS associé à un bulletin individuel est le libellé de profession renseigné par l'enquêté. Ce libellé est un texte qui constitue la réponse à une question parmi les trois suivantes :

— « *Quelle était votre profession principale ?* », dans le cas où l'enquêté ne travaille pas au moment où il répond au questionnaire mais a déjà travaillé ;
— « *Si vous n'êtes pas salarié, quelle est votre profession ?* », dans le cas où l'enquêté travaille au moment où il répond au questionnaire mais n'est pas salarié ;
— « *Quelle est votre profession principale ?* », dans le cas où l'enquêté est salarié au moment où il répond à l'enquête.

Plusieurs autres variables (catégorielles) sont à considérer pour déterminer la PCS associée à un bulletin individuel. Ces variables dépendent du type de profession du bulletin en question. La Table 1 donne les différentes variables annexes, leurs modalités et les types de profession pour lesquelles elles interviennent. Plusieurs variables annexes peuvent prendre de nombreuses valeurs différentes. On dénombre dans nos données 718 codes NAF2 différents pour l'activité des établissements et 177 catégories juridiques différentes (sans compter les valeurs manquantes).

La nouvelle version de la nomenclature PCS est associée à une liste de libellés (avec variables annexes associées) qui correspondent chacun à un unique code de la nomenclature. Par exemple, pour une profession actuelle salariée, le libellé de profession « courtier en assurances », avec une position professionnelle associée « ouvrier qualifié ou hautement qualifié, technicien d'atelier »,

		Profession		
Variable	Modalités	S	I	A
Situation principale	• Emploi • Apprentissage • Étude • Chômage • Retraite ou pré-retraite • Femme ou homme au foyer • Autre situation	✓	✓	✓
Statut professionnel actuel	• Travailleur indépendant ou à son compte • Chef d'entreprise salariée, PDG, gérant(e) minoritaire de SARL • Salarié(e) • Aide d'une personne dans son travail	✓	✓	
Statut professionnel antérieur	• Salarié(e) ou stagiaire rémunéré • Indépendant ou à votre compte • Aide d'une personne dans son travail			✓
Position professionnelle	• Manœuvre, ouvrier spécialisé • Ouvrier qualifié ou hautement qualifié, technicien d'atelier • Technicien (non cadre) • Agent de catégorie B de la Fonction Publique • Agent de maîtrise, maîtrise administrative ou commerciale, VRP • Agent de catégorie A de la Fonction Publique • Ingénieur, cadre d'entreprise • Agent de catégorie C ou D de la Fonction Publique • Employé (de bureau, de commerce, ...)	✓		
Nombre de salariés employés	• Aucun salarié employé • De 1 à 9 salariés employés • 10 salariés employés ou plus		✓	
Activité de l'établissement	Codes de la nomenclature d'activité NAF2	✓	✓	
Catégorie juridique de l'établissement	Codes de la nomenclature des catégories juridiques	✓	✓	
Effectifs de l'établissement	Modalités correspondant à des tranches d'effectifs	✓		

TABLE 1 – Variables annexes utilisées pour la codification de la PCS.

correspond au code PCS « 46D2 », soit « techniciens de la banque, des assurances et des organismes de sécurité sociale ». Pour une profession actuelle non-salariée, le libellé « énergéticien », lorsque l'enquêté travaille à son compte ou emploie moins de 10 personnes, correspond au code PCS « 22D6 », soit « indépendants d'autres prestations de service ». Au total, 5 762 libellés de profession, chacun décliné au masculin et au féminin, sont répertoriés dans la liste. Pour chaque enquête du recensement, une partie des bulletins individuels est donc codée automatiquement via l'*index numérique* (la liste de libellés).

Les enquêtes de recensement se font maintenant en partie sur Internet (60 % de réponse sur Internet et 40 % de réponse sur papier, proportions qui peuvent évoluer vers plus d'Internet dans le futur). Un moteur d'autocomplétion des libellés a été développé pour maximiser le nombre de bulletins individuels codés automatiquement via l'*index numérique* sur Internet. Lorsqu'un enquêté répondant sur Internet commence à renseigner son libellé de profession, des propositions de libellés pertinents s'affichent en fonction du texte qu'il a déjà entré. Les propositions évoluent à chaque fois que de nouveaux caractères sont ajoutés. À tout moment, l'enquêté peut choisir une des propositions si elle lui convient. On estime que le taux de bulletins codés via l'*index numérique* sur Internet est environ de 80 % (taux observé pour le pilote de la refonte de l'enquête Emploi). Une partie des bulletins papier est aussi codée automatiquement via l'*index numérique*, lorsque les enquêtés donnent spontanément un libellé de profession appartenant à l'index. On estime le taux de bulletins ainsi codés automatiquement à 35 % des bulletins papier. Cette estimation correspond au taux observé pour l'enquête annuelle de recensement (EAR) 2019.

Pour coder les bulletins non-codés automatiquement via l'*index numérique*, on choisit d'utiliser des méthodes d'apprentissage automatique, et plus précisément de classification supervisée. Cette approche a déjà été utilisée dans des contextes similaires [2]. Un classifieur est entraîné pour chaque tâche de codification (pour chaque type de profession). Les libellés restant à coder sont hors de l'*index numérique* et ont la forme de texte libre. Les classifieurs s'appuient sur un plongement lexical (c'est-à-dire une représentation vectorielle apprise) de ces libellés. La méthode est détaillée en section 2.

2 Méthodologie

2.1 Plongement lexical des libellés de profession

De nombreuses tâches en traitement du langage naturel nécessitent d'avoir des représentations vectorielles de documents (on appelle document une observation textuelle pour laquelle on souhaite effectuer une tâche donnée). C'est le cas pour la tâche de classification de document qui nous intéresse dans ce travail. Pour obtenir la représentation vectorielle d'un document, on passe souvent par la représentation vectorielle de chacun de ses mots. On peut ensuite choisir d'appliquer n'importe quelle méthode de classification aux documents « vectorisés ».

La représentation vectorielle la plus basique possible pour un mot au sein d'un corpus de documents est son encodage *one-hot* (ou encodage 1 parmi n) : soit un mot t_i indexé par $i \in [\![1\ ;\ N]\!]$ où N est le nombre de mots différents dans le corpus de documents considéré. Son encodage *one-hot* est

$$u_i = (0, ..., 0, 1, 0, ..., 0) \in [\![0\ ;\ 1]\!]^N,$$

un vecteur de 0 avec un 1 en $i^{\text{ème}}$ position. Pour faire de la classification de document, une première option possible est de considérer un modèle *bag of words*, où un document est considéré comme l'ensemble des mots qui le composent. Une représentation vectorielle possible d'un document est alors égale à la somme (ou à la moyenne) des encodages *one-hot* de ses mots. [3] met l'accent sur l'importance du pré-traitement appliqué aux documents en amont de cette

vectorisation pour une tâche de classification.

Des représentations vectorielles plus élaborées peuvent être dérivées d'un corpus de documents non-annoté : pour prendre un exemple qui reste très simple, la fréquence inverse de document est une mesure de l'importance d'un mot dans l'ensemble du corpus. Elle consiste à calculer le logarithme de l'inverse de la proportion de documents du corpus qui contiennent le mot, soit

$$\mathrm{idf}_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|},$$

où $|D|$ est le nombre total de documents dans le corpus et $|\{d_j : t_i \in d_j\}|$ est le nombre de documents où le terme t_i apparaît. Une représentation vectorielle possible pour t_i est alors

$$u_i = (0, ..., 0, \mathrm{idf}_i, 0, ..., 0) \in \mathbb{R}^N,$$

où le coefficient non-nul se trouve en $i^{\text{ème}}$ position. Avec le modèle *bag of words* décrit précédemment, la représentation vectorielle d'un document est son TF-IDF, introduit par [4] pour quantifier la pertinence de mots dans des recherches de documents.

Récemment, la communauté autour des *réseaux de neurones* a eu l'idée d'apprendre des représentations vectorielles de mots à l'aide de réseaux de neurones à propagation avant [5]. Un avantage de ces représentations vectorielles par rapport aux exemples simples donnés plus tôt est qu'elles sont denses et permettent ainsi d'éviter le fléau de la dimension (*curse of dimensionality*). En particulier, [6] construit des représentations vectorielles de mots à partir de corpus non-annotés avec les modèles de sacs de mots continus (CBOW : *continuous bag of words*) et *skip-gram*. Le modèle CBOW apprend des représentations vectorielles en cherchant à prédire un mot étant donné son contexte, c'est-à-dire étant donné les mots qui l'entourent (par exemple les deux mots à gauche et les deux mots à droite du mot à prédire). Le modèle *skip-gram* apprend des représentations vectorielles en essayant au contraire de prédire un contexte à partir d'un mot en entrée. D'autres modèles ont permis d'entraîner des ensembles de représentations vectorielles, utilisables pour n'importe quelle tâche en aval. C'est le cas par exemple de GloVe [7].

Il est possible d'apprendre des plongements lexicaux de manière supervisée, par exemple dans le cadre d'une tâche de classification de texte. C'est cette approche que nous avons choisie pour ce travail. Le classifieur utilisé repose au départ sur un modèle *bag of words* (BOW) pour la représentation des documents, couplé à un classifieur linéaire. Dans un modèle simple de *bag of words*, le document à classifier est représenté par l'ensemble des mots qui le composent. Le plongement lexical du document est alors (par exemple) égal à la moyenne des plongements lexicaux de chacun de ses mots. La fonction *softmax* f est utilisée en aval du classifieur linéaire pour calculer la distribution de probabilités sur l'ensemble des classes possibles. La matrice de plongement A et la matrice de coefficients du classifieur linéaire B sont apprises simultanément au cours de l'entraînement par descente de gradient, typiquement avec une fonction de perte de la forme

$$-\frac{1}{n}\sum_{i=1}^{n} y_i \log(f(Bg(A, x_i))),$$

où x_i constitue l'ensemble des mots du texte i, y_i sa vraie classe et g est une fonction qui donne le plongement d'un texte à partir d'une matrice de plongement. Notons que la matrice B peut être pré-entraînée, par exemple dans le cadre d'un apprentissage non-supervisé avec les modèles CBOW ou *skip-gram*.

L'amélioration principale de la méthode de plongement utilisée pour ce travail (introduite par [8]) par rapport à un modèle strictement *bag of words* est que les représentations vectorielles

des documents à classifier s'appuient à la fois sur leurs mots et sur des n-grammes. Un n-gramme est une sous-séquence de n éléments construite à partir d'une séquence donnée. En particulier, notre modèle s'appuie sur des n-grammes de caractères (des caractères consécutifs au sein du texte à classifier) et des n-grammes de mots (des mots consécutifs au sein du texte à classifier). Le modèle sur lequel sont construits les plongements lexicaux est ainsi un modèle *bag of n-grams*.

Utiliser un modèle *bag of n-grams* permet premièrement de tenir compte en partie de l'ordre des mots. Prenons l'exemple du libellé « **RESPONSABLE MAGASIN LOGISTIQUE** » : avec un modèle *bag of n-grams* on peut choisir d'inclure les plongements des bi-grammes de mots (« **RESPONSABLE MAGASIN** » et « **MAGASIN LOGISTIQUE** ») dans le calcul du plongement du libellé. On prend ainsi en compte le fait que les mots « **RESPONSABLE** » et « **LOGISTIQUE** » sont consécutifs, ce qui n'est pas le cas dans un modèle *bag of words*. Deuxièmement, le modèle *bag of words* ne permet pas non plus de vectoriser les mots s'ils n'ont pas été rencontrés au cours de l'apprentissage. L'utilisation de n-grammes de caractères permet d'obtenir un plongement même lorsqu'un mot ne figure pas dans le corpus d'apprentissage. Considérons par exemple le texte « **EMPLOI** » , dans le cas où les seuls n-grammes pris en compte sont les tri-grammes de caractères. Ce texte est découpé dans notre modèle en l'ensemble de jetons suivant : <**EM**, **EMP**, **MPL**, **PLO**, **LOI**, **OI**> en plus de <**EMPLOI**>. Les caractères < et > symbolisent le début et la fin d'un mot. Ils permettent de distinguer un n-gramme provenant d'un mot et un mot de n lettres. Même si le plongement du jeton <**EMPLOI**> n'a pas été appris au cours de l'entraînement, une représentation vectorielle peut être calculée à partir des plongements des tri-grammes. A priori si certains mots proches sémantiquement du mot « **EMPLOI** » figuraient dans le corpus d'entraînement, la représentation vectorielle de ces mots sera proche de celle calculée pour « **EMPLOI** » à partir de ses tri-grammes. Naturellement, la représentation en n-grammes est assez robuste aux fautes d'orthographe ou de frappe, etc. [8] indique que la méthode utilisée ici donne pour une tâche de classification de document des résultats comparables à des modèles beaucoup plus complexes (comme BERT [9] ou le modèle dérivé CamemBERT [10] pour la langue française, qui ont des performances *state-of-the-art* pour de nombreuses tâches de traitement du langage), tout en étant beaucoup plus rapide. Notons qu'un modèle *bag of n-grams* est aussi utilisable pour calculer des représentations vectorielles de manière non-supervisée [11].

Notre modèle est paramétré par la taille minimale et la taille maximale des n-grammes de caractères, ainsi que par la taille maximale des n-grammes de mots à inclure dans le calcul du plongement lexical des libellés. Ce plongement est égal à la moyenne des plongements des jetons (mots, n-grammes de mots et de caractères) définis à partir du paramétrage.

2.2 Prise en compte des variables annexes

Les variables annexes qui peuvent intervenir dans la codification de la PCS sont détaillées dans la Table 1. Certaines variables ont beaucoup de modalités et il est donc important d'appliquer une méthode de réduction de dimension aux variables annexes en amont du classifieur linéaire.

Une première possibilité pour réduire la dimension des variables annexes est de les concaténer au libellé de profession pour n'avoir en entrée du classifieur que le plongement d'un libellé « enrichi ». Par exemple, considérons un bulletin individuel avec le libellé de profession « **RESPONSABLE MAGASIN LOGISTIQUE** » , la situation principale déclarée « **Emploi** » (modalité 1) et la position professionnelle déclarée « **Employé (par exemple : de bureau, de commerce, de la restauration, de maison)** » (modalité 9). Le libellé enrichi sera dans ce cas « **RESPONSABLE MAGASIN LOGISTIQUE SITUATION_1 POSITION_9** ». On

ajoute des préfixes avec le nom de variable pour chaque variable annexe afin qu'un plongement différent soit bien associé à une modalité identique pour deux variables annexes différentes.

Cette méthode semble donner des résultats satisfaisants en pratique mais présente des défauts. Par exemple, des n-grammes de caractères issus de variables annexes sont pris en compte dans le calcul du plongement des libellés enrichis, ce qui n'est pas pertinent.

Un modèle plus logique consiste à entraîner, en plus de la matrice A de plongement des libellés, une matrice de plongement indépendante M_j pour chaque variable annexe, avec $j \in [\![1\ ;\ C]\!]$ où C est le nombre de variables annexes : chaque modalité de chaque variable annexe a alors une représentation vectorielle. La dimension de l'espace de plongement D_j est un hyperparamètre à fixer pour chaque variable. Pour une observation i, la représentation vectorielle dense u_i associée au libellé de profession est la moyenne des représentations vectorielles des jetons qui le composent (mots, n-grammes de mots et n-grammes de caractères). On a également une représentation vectorielle dense v_{ij} associée à chacune des C variables catégorielles. Pour tout $j \in [\![1\ ;\ C]\!]$:

$$v_{ij} = (M_j)_{z_{ij}},$$

c'est-à-dire que cette représentation est égale à la ligne de la matrice M_j correspondant à la modalité z_{ij} prise par la $j^{\text{è}}$ variable catégorielle. La représentation vectorielle dense de l'observation est

$$w_i = (u_i, v_{i1}, ..., v_{iC}),$$

la concaténation des représentations vectorielles de toutes les variables.

2.3 Classifieur

Un classifieur linéaire prend la représentation vectorielle dense des observations en entrée pour prédire la PCS. Plusieurs options sont possibles pour l'algorithme d'entraînement. La première option a déjà été brièvement mentionnée en 2.1. Soit f la fonction *softmax*, c'est-à-dire la fonction qui à un vecteur z de $\mathbb{R}^n$ associe

$$f(z) = \left[\frac{\exp(z_1)}{\sum \exp(z_i)}, ..., \frac{\exp(z_n)}{\sum \exp(z_i)}\right] \in \mathbb{R}^n.$$

Les probabilités pour chaque classe prédites par le classifieur pour une observation i sont égales à $f(Bw_i)$, où w_i est la représentation vectorielle dense de l'observation i et B est la matrice de coefficients du classifieur linéaire, de dimension $K \times (D + \sum_j D_j)$, avec K le nombre de classes. L'apprentissage de la matrice B et de toutes les matrices de plongement se fait par descente de gradient. La fonction de perte utilisée est la *cross-entropy* :

$$L(w_1, ...w_n) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(f(Bw_i)).$$

Les matrices de plongement A et M_j pour $j \in [\![1\ ;\ C]\!]$ sont entraînées en même temps que la matrice B.

La seconde option pour l'algorithme d'entraînement correspond à un schéma *one-versus-all*. Dans ce schéma, un classifieur binaire indépendant est entraîné pour chaque classe. Cette variation est particulièrement utile lorsqu'on souhaite pouvoir prédire plusieurs classes pour un même libellé. Ce n'est pas le cas ici puisqu'un seul code PCS correspond à un bulletin individuel donné, mais en choisissant systématiquement le code correspondant à la probabilité la plus élevée, le schéma *one-versus-all* donne de meilleurs résultats que l'option reposant sur le *softmax*. C'est donc la stratégie *one-versus-all* qui a été retenue pour notre modèle.

L'apprentissage supervisé nécessite une base de données dite *d'apprentissage* pour apprendre les paramètres du modèle en optimisant la fonction de perte. Les détails sur l'échantillonnage de la base d'apprentissage sont donnés en section 3.

2.4 Indice de confiance

Le moteur de codification Sicore utilisé dans le passé pour la codification automatique de la PCS est un ensemble de règles de décision qui ne couvre pas tous les libellés (et variables annexes associées) possibles. Sicore n'est ainsi pas capable de coder tous les bulletins individuels. Pour chaque enquête du recensement, une partie des bulletins était déclarée « incodable » par Sicore et était envoyée en reprise manuelle.

Au contraire de Sicore, les classifieurs décrits dans cette étude prédisent des probabilités d'appartenance à chaque classe de la nomenclature PCS pour n'importe quel bulletin individuel donné en entrée. Le code retenu est naturellement le code associé à la probabilité la plus élevée. Plus cette probabilité est élevée, plus le classifieur est confiant sur la codification. Un indice de confiance qui est aussi régulièrement utilisé est la différence entre la probabilité prédite pour la classe la plus probable et la probabilité prédite pour la seconde classe la plus probable. Cet indice de confiance est compris entre 0 (confiance nulle) et 1 (confiance totale).

En théorie, la reprise manuelle pour une campagne du recensement pourrait être modulée en fonction des indices de confiance du classifieur sur les bulletins de l'échantillon à coder. Si le classifieur se trompe souvent sur les bulletins pour lesquels il affiche un indice de confiance faible, il est naturel d'envoyer ces bulletins en reprise manuelle, au contraire des bulletins pour lesquels l'indice de confiance est élevé (si en effet pour ces bulletins la prédiction du classifieur s'avère souvent correcte). Cette question sera abordée dans la section 4.

3 Campagne d'annotation

Les méthodes d'apprentissage supervisé développées en 2 nécessitent des données annotées pour apprendre les paramètres du modèle en optimisant la fonction de perte. Suite au report de l'enquête de recensement de 2021, des ressources ont été débloquées pour une campagne de labellisation avec pour but d'annoter un échantillon de bulletins individuels dans la nouvelle nomenclature PCS. Les modalités de l'échantillonage des bulletins à annoter sont décrites dans les sous-sections suivantes.

3.1 Taille des échantillons d'entraînement

La taille des échantillons d'entraînement (un par type de profession) à annoter au cours de la campagne d'annotation a été déterminée à partir de tests préliminaires sur des données de bulletins individuels annotés dans la nomenclature PCS 2003. La méthode de codification utilisée pour ces tests a été décrite en 2 [1]. Ici on a choisi de concaténer les variables annexes au libellé de profession pour n'avoir en entrée qu'un libellé enrichi. On a aussi choisi pour ces tests d'entraîner C classifieurs binaires en aval de la phase de plongement lexical (schéma *one-versus-all*).

1. C'est d'ailleurs également à partir de tests préliminaires sur les données annotées en PCS 2003 que cette méthode a été identifiée comme particulièrement pertinente pour une tâche de codification dans une nomenclature comme celle de la PCS. Pour les tests l'implémentation proposée par la librarie open-source fastText a été utilisée

Les tailles des bases d'apprentissage minimales pour répondre aux contraintes de qualité (avoir une qualité équivalente à la qualité de codification de Sicore, au niveau de la part de codification automatique et de sa précision) ont été estimées à :

— 70 000 bulletins pour la profession actuelle salariée ;
— 10 000 bulletins pour la profession actuelle non-salariée ;
— 5 000 bulletins pour la profession antérieure (codée sur 2 positions).

Les tailles finales fixées suite aux tests pour la campagne d'annotation sont légèrement supérieures, pour avoir la possibilité d'améliorer la performance de la codification automatique par rapport à celle de Sicore. Ainsi, 110 000 bulletins individuels ont été annotés :

— 90 000 bulletins pour la profession actuelle salariée ;
— 12 500 bulletins pour la profession actuelle non-salariée ;
— 7 500 bulletins pour la profession antérieure.

Les méthodes de classification supervisée avec entraînement sur une base d'apprentissage, imposent que cette dernière soit de très bonne qualité. C'est pourquoi les données d'entraînement ont été annotées par triple codage : deux annotateurs proposent chacun un code à l'aveugle. Si les codes sont identiques, l'annotation est terminée. En cas de divergence des deux premiers codes, un arbitre détermine le code finalement retenu (qui peut être différent des codes proposés par les deux premiers annotateurs).

3.2 Stratégie de sélection des échantillons d'entraînement

Les tests conduits sur les données annotées en PCS 2003 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, avant le tirage des échantillons, on a d'abord procédé à une suppression des doublons (libellé de profession et variables annexes) afin d'éviter de faire annoter deux fois le même bulletin.

L'approche la plus simple consisterait à tirer les 110 000 bulletins individuels via un sondage aléatoire simple à probabilités d'inclusion égales. Plusieurs alternatives ont été étudiées.

3.2.1 Tirage aléatoire avec probabilité d'inclusion proportionnelle à la fréquence d'apparition des caractéristiques individuelles

La première alternative 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). Les variables utilisées pour calculer les probabilités d'inclusion sont celles qui ont été identifiées comme étant les plus porteuses d'information pour coder dans la nomenclature des professions. Elles sont listées dans la Table 2.

Soit un échantillon de test (w_i, y_i) pour $i \in [\![1 \ ; \ n]\!]$ tiré aléatoirement. La précision d'un classifieur $\hat{h}$ sur cet échantillon de test est égale à

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\hat{h}(w_i) = y_i).$$

C'est cette métrique qui a été retenue pour choisir la stratégie optimale d'échantillonnage à volume donné. Les précisions des classifieurs, pour chaque type de profession, en fonction de la taille des échantillons d'entraînement pour l'approche de base (sondage aléatoire simple) et l'alternative 1 sont données en Figure 1. L'alternative 1 apparaît comme meilleure que l'approche

Profession	Variables de tirage
Profession actuelle salariée	• Libellé de profession • Statut professionnel • Position professionnelle • Secteur d'activité de l'établissement employeur en NAF5
Profession actuelle non-salariée	• Libellé de profession • Statut professionnel • Pluralité de salariés employés • Tranche d'effectif de l'entreprise • Secteur d'activité de l'établissement employeur en NAF5
Profession antérieure	• Libellé de profession • Statut professionnel

TABLE 2 – Variables utilisées pour le tirage avec probabilités d'inclusion proportionnelles au nombre d'occurrences.

de base : pour tous les types de profession, le tirage systématique améliore la précision des modèles.

3.2.2 Tirage aléatoire à partir de plongements lexicaux

L'alternative 1 s'appuie sur les variables brutes issues des bulletins individuels. 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 plongements lexicaux pour projeter les libellés dans un espace vectoriel de dimension limitée. Deux mots ou textes sémantiquement proches sont représentés dans cet espace par deux vecteurs numériquement proches. Des méthodes de clustering (K-means, CAH, DBSCAN...) peuvent être utilisées dans l'espace de plongement pour constituer n groupes distincts. Les individus au sein d'un groupe doivent être aussi homogènes que possible. On peut ensuite tirer un représentant selon une loi uniforme dans chaque groupe.

Plusieurs variantes de cette alternative (plongements différents, méthodes de clustering différentes) ont été testées mais l'approche n'a finalement pas été retenue, car elle ne présentait pas de vrai gain de performance par rapport à l'alternative 1 plus simple.

3.2.3 Tirage avec une stratégie d'*active learning* : tirage d'un petit échantillon puis sélection de nouvelles informations 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 classifieur 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 d'un classifieur entraîné avec ces observations supplémentaires [12].

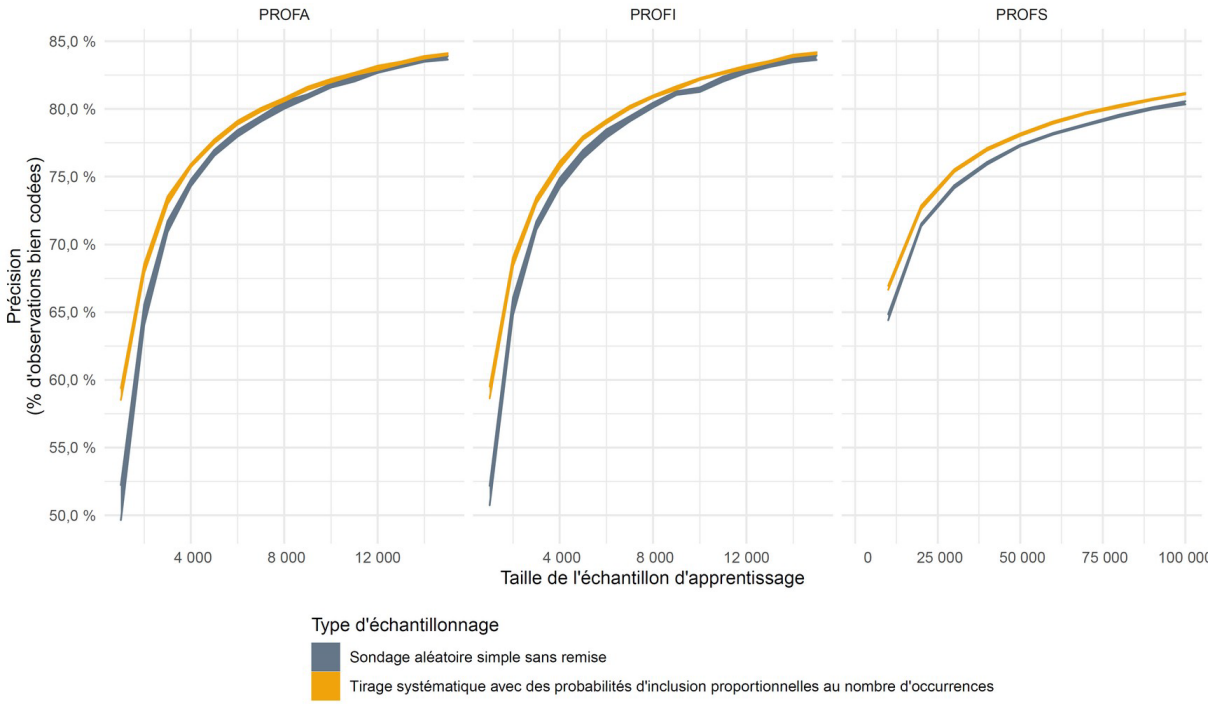


FIGURE 1 – Comparaison de l'efficacité de l'échantillonnage proportionnel à la taille (alternative 1) et du sondage aléatoire simple avec probabilités d'inclusion égales.

En pratique, un premier échantillon de petite taille n_{initial} est tiré (par sondage aléatoire simple ou tirage systématique sur les variables initiales par exemple). Une fois celui-ci annoté, un premier classifieur est entraîné. Puis, une prédiction à partir de ce classifieur est réalisée sur l'ensemble des données de la base de sondage excepté l'échantillon déjà annoté. 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.

Au travers des expérimentations menées, il apparaît que la stratégie d'*active learning* est meilleure pour des volumes de données plus grands (voir Figure 2) lorsque l'échantillon initial est tiré selon un sondage aléatoire simple et avec les paramètres donnés dans la Table 3. 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 n'étions pas en mesure de mettre en oeuvre ce scenario avec un pas trop petit car cela demandait au sein de l'application de labellisation de recalculer les modèles très régulièrement. L'intégration de cette fonctionnalité soulevait des questions délicates d'implémentation qui ne pouvaient pas être résolues dans un délai court (risque fort de latences par exemple).

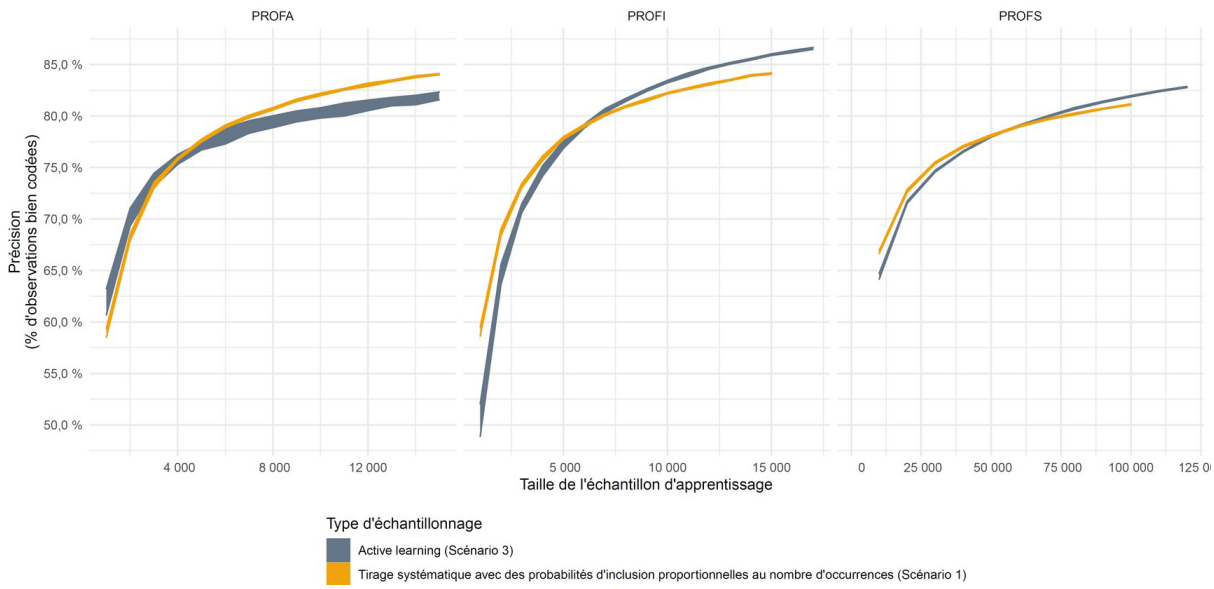


FIGURE 2 – Comparaison de l'efficacité de l'échantillonnage avec stratégie d'*active learning* (alternative 3) et du sondage aléatoire avec probabilité d'inclusion proportionnelle à la fréquence d'apparition (alternative 1).

Profession	Taille de l'échantillon initial n_{initial}	Pas n_{pas}
Profession antérieure	1 000	1 000
Profession actuelle non-salariée	1 000	1 000
Profession actuelle salariée	10 000	10 000

TABLE 3 – Paramètres choisis pour la stratégie d'*active learning*.

3.2.4 Tirage d'un échantillon principal puis d'une réserve déterminée par *active learning*

Malgré les difficultés d'intégration de l'alternative précédente, il est possible de l'adapter pour tenter de profiter des avantages attendus sur la précision des modèles. Un échantillon principal d'une taille $n_{\text{principal}}$ est au préalable sélectionné par tirage systématique (alternative 1). 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.

D'après la Figure 3, où la taille des échantillons principaux ont été fixés à 5 000, 10 000 et 70 000 observations respectivement pour la profession antérieure, la profession actuelle non-salariée et la profession actuelle salariée, cette stratégie gagne en performance avec le volume de données. Elle ne dépasse le scénario 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.

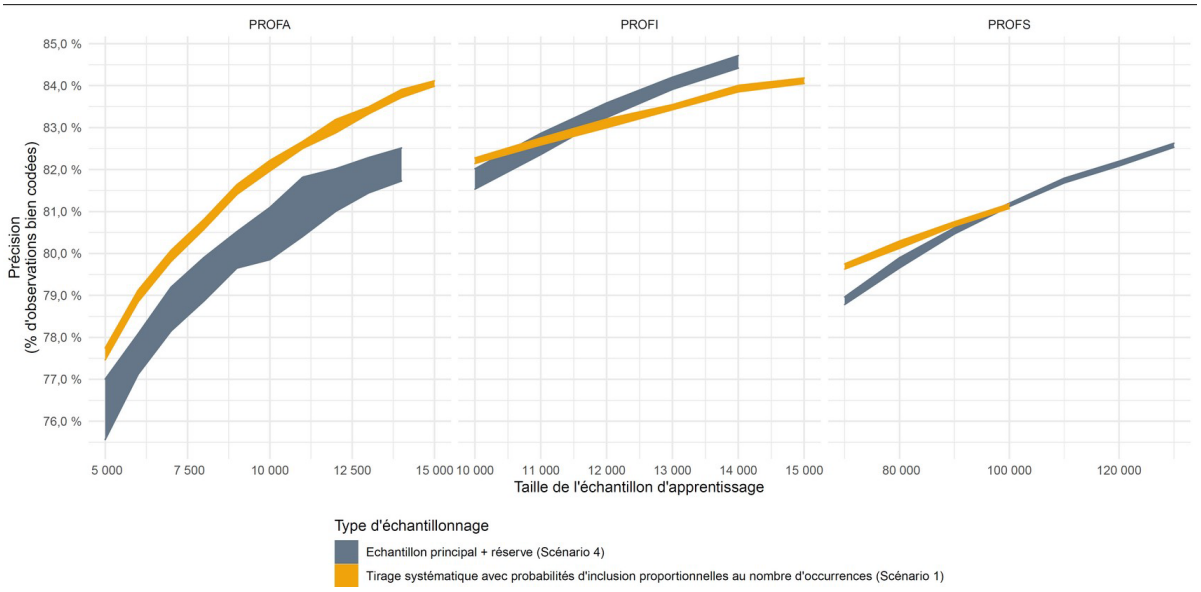


FIGURE 3 – Comparaison de l'efficacité de l'échantillonnage avec stratégie d'échantillon principal et d'échantillon de réserve (alternative 4) et du sondage aléatoire avec probabilité d'inclusion proportionnelle à la fréquence d'apparition (alternative 1).

3.3 Déroulement de la campagne d'annotation

Le département de la démographie et le service recensement national (SeRN) ont organisé cette opération en lien avec le SSP Lab, la division recueil et traitement de l'information et le pôle PCS de Besançon. En estimant le volume de bulletins traités en moyenne par une personne et par jour à 110, comme dans le cadre des campagnes Recap, la charge totale a été estimée à environ 55 ETP sur 3 mois. Chaque gestionnaire a reçu, au préalable, une formation pour coder dans cette nouvelle nomenclature. Celle-ci était délivrée par les experts du pôle PCS. Elle comprenait une première journée théorique et une demi-journée supplémentaire pour balayer des cas pratiques. Les travaux de codification se sont ensuite étalés du 15 février au 17 mai 2021. Avant d'accéder à la codification des près de 120 000 bulletins individuels, les annotateurs devaient coder au préalable un étalon-or de 20 observations construit par le pôle PCS pour s'assurer que les consignes données lors des formations étaient bien assimilées. En moyenne, 13 professions sur 20 étaient codées parfaitement au niveau le plus fin. Ce taux de 65 % de professions bien codées pour les bulletins repris manuellement correspond à celui mesuré au travers des campagnes Qualité sur la PCS 2003.

Une application ad-hoc a été conçue pour faciliter l'annotation, veiller à ce que les différents codeurs soient tous différents lors du triple codage, rassembler les données ou encore simplifier le suivi. Pour l'annotateur, lors du traitement de chacune des tâches, les caractéristiques portant sur la profession étaient affichées sur la moitié gauche de l'écran. L'autre moitié permettait de réaliser l'annotation, c'est à dire de sélectionner un code PCS 2020. Si les informations disponibles ne permettaient pas de coder au niveau le plus fin de la nomenclature (4 positions pour la profession actuelle et 2 positions pour la profession antérieure), il était possible de choisir un code à un niveau plus agrégé, voire de coder à blanc. Le choix du code pouvait s'effectuer de plusieurs manières :

— à partir d'un système de recommandation entraîné sur les EAR 2015 à 2019 en PCS 2003 et d'une table de passage PCS 2003 vers PCS 2020 ;
— grâce à un système d'autocomplétion pour rapprocher le libellé d'origine d'un libellé contenu dans l'index des professions de la nomenclature PCS 2020 et ainsi obtenir le

code « officiel » ;
— manuellement au travers de la nomenclature.

Le recours à la première méthode devait être privilégié uniquement pour les cas les plus simples afin de rendre la codification plus rapide. Pour les tâches plus complexes, l'utilisation de l'index des professions était fortement recommandée lors des formations. Un usage abusif du modèle appris sur l'ancienne nomenclature principalement au travers des environnements Sicore aurait engendré trop d'erreurs. En effet, bien qu'ergonomique, le système de recommandation introduit possiblement du *satisficing*.

FIGURE 4 – Capture d'écran de l'interface d'annotation utilisée pendant la campagne

4 Résultats

4.1 Statistiques descriptives

4.1.1 Profession actuelle salariée

Le jeu d'entraînement pour la profession actuelle salariée est finalement composé de données issues de 90 000 bulletins individuels, dont 70 000 constituent un jeu initial et les 20 000 restants un jeu de réserve qui a été échantillonné avec une stratégie d'*active learning* en une itération, comme décrit en 3.2.4. Dans le jeu d'entraînement, on compte 428 codes PCS différents (en considérant la modalité « non-codable » comme un code). C'est plus que le nombre de professions qui existent au niveau le plus fin de la nomenclature, ce qui est normal étant donné que les annotateurs pouvaient choisir un code à un niveau supérieur si besoin. Les libellés de profession bruts du jeu d'entraînement sont propres et un nettoyage standard leur est appliqué : on retire la ponctuation et les mots vides de sens. Les libellés nettoyés sont courts, avec 2,4 mots en moyenne par libellé et un écart interdécile (différence entre le neuvième et le premier décile) de 3 mots. Seulement 1.2 % des libellés ont 6 mots ou plus et le libellé le plus long de l'échantillon contient 12 mots.

Les 20 mots les plus fréquents au sein des libellés sont « agent », « responsable », « chef », « technicien », « assistante », « service », « ingénieur », « technique », « adjoint », « directeur », « commercial », « employé », « chargé », « gestionnaire », « classe », « A », « ouvrier », « administrative », « production » et « catégorie ». Les bi-grammes de mots les plus fréquents sont « agent entretien », « chef projet », « responsable service », « adjoint technique », « fonction publique », « assistante direction », « agent entretien », « technicien maintenance », « agent maîtrise » et « agent technique ».

La très grande majorité des observations ont la situation principale « Emploi », et seule une très faible part des bulletins ont pour situation principale « Chômage » ou « Femme ou homme au foyer » (0,8 %), ce qui est attendu. La variable n'est pas renseignée pour 0,6 % des bulletins.

Plus de 95 % des observations ont pour statut professionnel la modalité « Salariée ». Par ailleurs, cette variable n'est pas renseignée pour 3 % des bulletins et environ 1 % des observations a la modalité « Chef d'entreprise salarié(e), PDG, gérant(e) minoritaire de SARL ». Il ne reste donc qu'une très faible part des bulletins (environ 0,6 %) avec la modalité « Travailleur indépendant ou à son compte », ce qui est rassurant : ces bulletins devraient plutôt être associés à une profession actuelle non-salariée.

La distribution de la variable de position professionnelle au sein du jeu d'entraînement pour la profession actuelle salariée est donnée en Figure 5. Environ un quart des observations correspondent à une profession d'« employé (de bureau, de commerce, ...) ». Les positions d'« ingénieur ou cadre d'entreprise » sont aussi fortement représentées, à hauteur de 18 % des bulletins. Les autres positions professionnelles ont des taux de prévalence compris entre 4 % et 11 %. La variable n'est pas renseignée dans 4 % des cas. Cette variable devrait intuitivement être discriminante dans la détermination de la PCS. Cette intuition est confirmée par la Figure 6. La distribution du groupe socioprofessionnel (premier niveau de la nomenclature PCS) varie énormément en fonction de la position professionnelle renseignée sur le bulletin individuel. Par exemple, le code PCS appartient en majorité au groupe socioprofessionnel « Ouvriers » lorsque la position professionnelle renseignée est « Manoeuvre, ouvrier spécialisé » ou « Ouvrier qualifié ou hautement qualifié, technicien d'atelier ». De même, le code PCS appartient très souvent au groupe socioprofessionnel « Professions intermédiaires » lorsque la position professionnelle est « Agent de maîtrise, maîtrise administrative ou commerciale, VRP », etc. On remarque que, conformément à ce qui est attendu, très peu d'observations ont un code du groupe socioprofessionnel « Artisans, commerçants et chefs d'entreprise ». Ce groupe correspond davantage aux professions non-salariées.

L'activité de l'entreprise est une variable qui est aussi *a priori* importante pour déterminer la PCS. Dans le jeu d'entraînement utilisé pour la profession actuelle salariée, on dénombre 713 modalités différentes (et donc 713 codes NAF différents) en plus de l'activité non renseignée (qui représente 4% des cas). Seulement 9 activités ont un taux de prévalence supérieur à 1 %, et représentent ensemble environ 18 % des bulletins :

- Administration publique générale ;
- Activités hospitalières ;
- Enseignement secondaire général ;
- Autres intermédiations monétaires ;
- Ingénierie, études techniques ;
- Enseignement supérieur ;
- Défense ;
- Autres organisations fonctionnant par adhésion volontaire ;
- Institution de retraite supplémentaire.

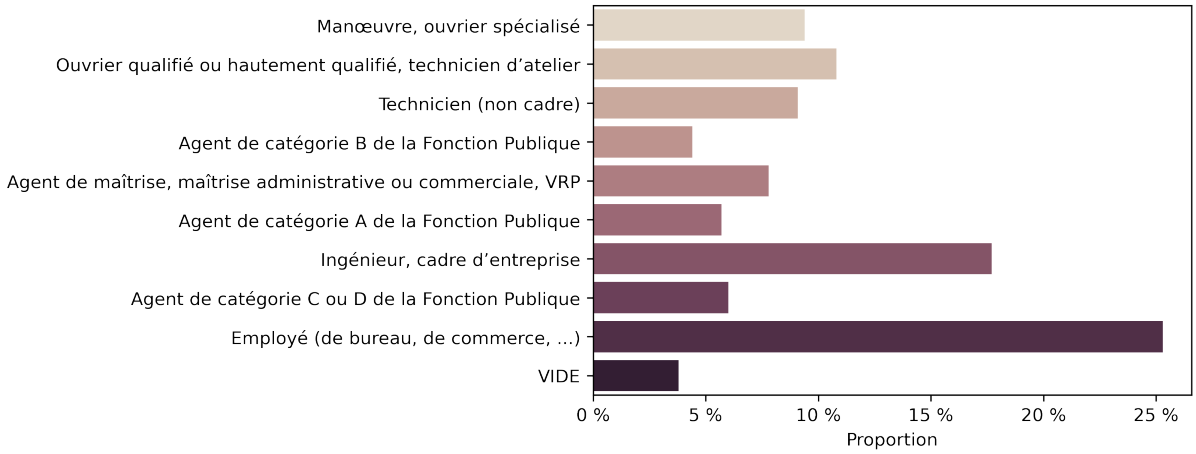


FIGURE 5 – Distribution des positions professionnelles pour le jeu d'apprentissage de profession actuelle salariée.

FIGURE 6 – Distribution du groupe socioprofessionnel en fonction de la position professionnelle pour le jeu d'apprentissage de profession actuelle salariée.

Les 231 activités avec un taux de prévalence supérieur à 0,1 % représentent 81 % du nombre total d'observations. Au vu du très grand nombre de modalités et de la nomenclature utilisée, il est envisageable de se placer à un niveau plus élevé de la nomenclature et de ne conserver que les 2 ou 3 premiers caractères.

La variable de tranche d'effectifs n'est pas très bien renseignée : elle est manquante dans 13 % des cas. Dans 4 % des cas, la variable n'est pas manquante mais l'effectif de l'établissement employeur n'est pas connu. L'établissement employeur compte moins de 10 salariés dans 12 % des cas, entre 10 et 99 salariés dans 22 % des cas, entre 100 et 999 salariés dans 24 % des cas et plus de 1 000 salariés dans 25 % des cas.

4.1.2 Profession actuelle non-salariée

La profession non-salariée est *a priori* plus facile à coder que la profession salariée : on compte 349 codes PCS différents (« non-codables » inclus) dans l'échantillon d'entraînement. Une grande partie des codes PCS associés à des professions non-salariées est supposée appartenir au groupe socioprofessionnel « Artisans, commerçants et chefs d'entreprise ». C'est pour cela que la taille du jeu d'entraînement à échantillonner qui a été fixée pour la profession actuelle non-salariée est plus faible que pour la profession actuelle salariée. Finalement le jeu d'entraînement contient 12 500 observations, dont 10 000 constituent le jeu initial et les 2 500 restants le jeu de réserve.

Comme pour la profession salariée, les libellés de profession sont courts, avec en moyenne 1,9 mot par libellé. L'écart interdécile est de 2 mots, 0,3 % des libellés ont 6 mots ou plus et le libellé le plus long contient 10 mots. Les mots les plus fréquents sont naturellement différents des mots les plus fréquents pour les libellés de profession salariée : « gérant », « commerçant », « chef », « artisan », « directeur », « agent », « gérante », « commercial », « entreprise », « consultant », « entrepreneur », « entreprise », « société », « ingénieur », « A », « auto », « médecin », « général », « président », « vendeur ». De même les bi-grammes de mots les plus fréquents sont « chef entreprise », « directeur général », « auto entrepreneur », « gérant société » et « agent commercial ».

La situation principale « Emploi » est encore fortement majoritaire et concerne 80 % des observations. Une part significative des bulletins est associée aux situations « Retraite ou préretraite » (4 %) et « Autre situation » (11 %). Concernant le statut professionnel, presque 9 observations sur 10 ont pour modalité « Travailleur indépendant ou à son compte » (62 %) ou « Chef d'entreprise salariée, PDG, gérant(e) minoritaire de SARL » (27 %). Les modalités « Salarié(e) » et « Aide d'une personne dans son travail » concernent respectivement 6 % et 2 % des observations. La part non-négligeable associée à la modalité « Salarié(e) » correspond possiblement à des erreurs à un niveau du questionnaire. Enfin, la variable n'est pas renseignée dans 4% des cas.

Pour la profession non-salariée, le questionnaire demande le nombre de salariés employés, avec 3 modalités possibles. Au sein du jeu d'entraînement, 54 % des observations concernent des enquêtés qui n'emploient aucun salarié, 2 5% des enquêtés qui emploient 1 à 9 salariés et 10 % 10 salariés ou plus. Dans les autres cas (11 % des observations), la variable n'est pas renseignée.

La Figure 7 donne la distribution des groupes socioprofessionnels au sein du jeu d'apprentissage pour la profession actuelle non-salariée. Comme attendu, une forte proportion des observations (58 %) a un code PCS du groupe « Artisans, commerçants et chefs d'entreprise ». Les groupes « Cadres et professions intellectuelles supérieures » et « Professions intermédiaires » sont aussi fortement représentés, avec respectivement 17 % et 11 %. En outre, 5 % des observations

n'ont pas pu être codées par les annotateurs.

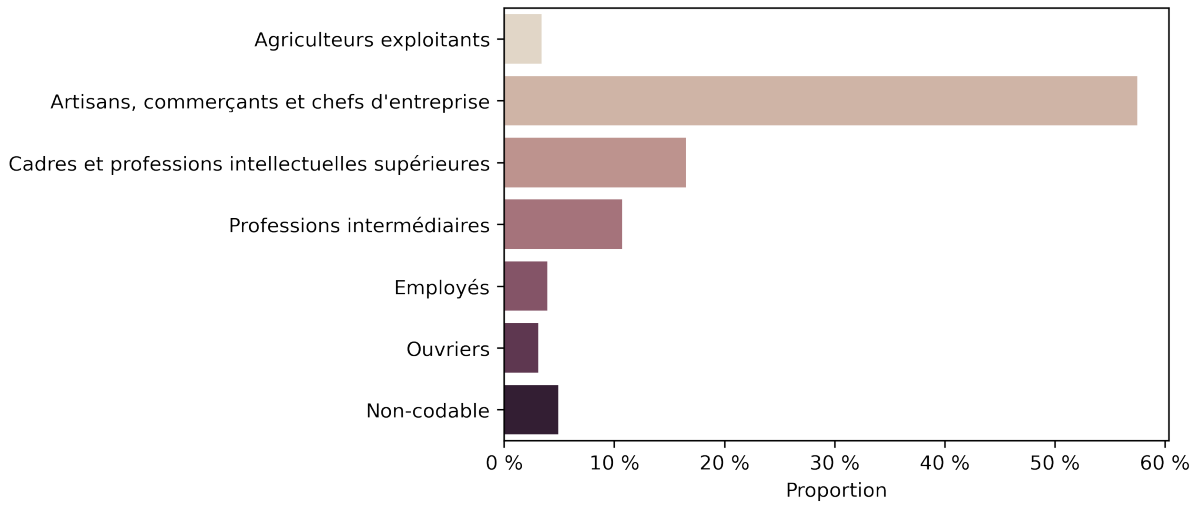


FIGURE 7 – Distribution des groupes socioprofessionnels pour le jeu d'apprentissage de profession actuelle non-salariée.

Si les activités les plus fréquentes ne sont pas les mêmes que pour la profession salariée, la distribution des activités des établissements est similaire pour les deux types de professions actuelles : 13 codes d'activité différents ont un taux de prévalence supérieur à 1 % (pour un total de 19 % des observations). Les 209 activités avec un taux de prévalence supérieur à 0,1 % représentent 80% du nombre total d'observation. L'activité n'est pas renseignée dans 7 % des cas.

4.1.3 Profession antérieure

La profession antérieure n'est codée que sur les deux niveaux les plus élevés de la nomenclature PCS. On ne compte ainsi dans l'échantillon d'entraînement que 35 classes différentes (« non-codables » inclus). Comme pour la profession actuelle non-salariée, le jeu d'apprentissage pour la profession antérieure est de taille réduite. Il contient 7 500 observations : 6 000 pour le jeu de données initial et 1 500 pour le jeu de réserve. Les libellés de profession sont courts, de 2.4 mots en moyenne. Seulement 1% des libellés comporte 6 mots ou plus. On retrouve dans les mots les plus fréquents des mots fréquents pour les professions actuelles : « agent », « responsable », « employé », « chef », « employée », « cadre », « ouvrier », « technicien », « secrétaire », « directeur » sont les mots les plus souvent rencontrés.

Seulement une faible part des observations (3 %) a pour situation principale la modalité « Emploi », ce qui est rassurant car ces observations ne devraient en théorie pas correspondre à une profession antérieure. La modalité la plus fréquente est de loin « Retraite ou pré-retraite » qui représente 59 % des observations. La modalité « Chômage » est aussi renseignée pour un nombre important d'observations (16 %). La grande majorité des professions antérieures étaient des professions salariées (ou de stagiaire rémunéré), avec un taux de prévalence de 82 %, et 9 % étaient des professions non-salariées. On compte en plus 2 % d'aide d'une autre personne dans son travail sans rémunération. La variable de statut antérieur n'est par ailleurs pas renseigné dans les 7 % de cas restants.

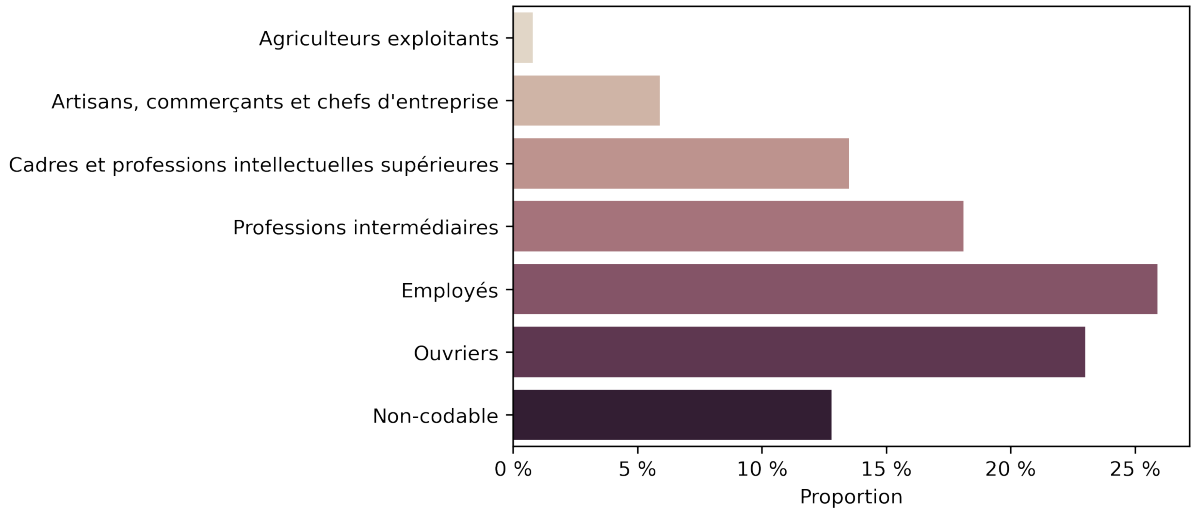


FIGURE 8 – Distribution des groupes socioprofessionnels pour le jeu d'apprentissage de profession antérieure.

La Figure 8 donne la distribution des groupes socioprofessionnels pour la profession antérieure. On observe une part élevée de codes dans les groupes « Professions intermédiaires », « Ouvriers » et « Employés » (allant de 18% à 26%). En outre, beaucoup plus d'observations sont « non-codables » (13%) que dans le cas de la profession actuelle.

4.2 Performance des classifieurs

Les classifieurs décrits en section 2 sont entraînés sur les données d'entraînement présentées dans la sous-section précédente. Leur performance est évaluée sur des jeux de test, tirés indépendamment des échantillons d'apprentissage selon un tirage systématique où les probabilités d'inclusion sont proportionnelles au nombre d'occurrences dans la base d'échantillonnage. Pour évaluer la performance d'un classifieur $\hat{h}$, la proportion d'observations bien classées doit être estimée en pondérant par l'inverse des probabilités d'inclusion, ce qui conduit à l'estimateur approximativement sans biais suivant :

$$\frac{\sum_{i=1}^{n} \frac{t_i \mathbb{1}(\hat{h}(w_i) = y_i)}{\pi_i}}{\sum_{i=1}^{n} \frac{t_i}{\pi_i}},$$

où l'indicatrice $\mathbb{1}(\hat{h}(w_i) = y_i)$ vaut 1 lorsque le classifieur prédit bien la valeur attendue pour l'observation (w_i, y_i), t_i est le nombre d'occurrences de l'observation dans la base d'échantillonnage (l'EAR 2020) et π_i est la probabilité d'inclusion de l'observation i dans l'échantillon de test.

Un échantillon de test a été tiré pour chaque type de profession : il compte 5 000 observations pour la profession actuelle salariée, 2 000 observations pour la profession actuelle non-salariée et 2 000 observations pour la profession antérieure.

Pour les classifieurs qui intègrent les variables annexes à travers les libellés enrichis, les hyperparamètres choisis pour chaque type de profession sont donnés en Table 4. Ces hyperparamètres ont été fixés par *grid search*, en utilisant un jeu de validation de taille réduite. Pour les classifieurs qui intègrent les variables annexes en marge du libellé de profession, les hyperparamètres

sont donnés en Table 5. Les proportions pondérées d'observations bien classées sont données en Table 6. On observe des performances similaires pour les deux modèles.

	PROFS	PROFI	PROFA
Dimension de l'espace de plongement	150	100	100
Taille maximale des n-grammes de mots inclus dans le calcul du plongement	3	3	3
Taille maximale des n-grammes de caractères inclus dans le calcul du plongement	4	5	5
Taille minimale des n-grammes de caractères inclus dans le calcul du plongement	3	3	3

TABLE 4 – Hyperparamètres choisis pour les classifieurs avec plongement des libellés enrichis.

4.3 Indices de confiance et reprise

On peut décider d'introduire de la reprise manuelle pour un sous-ensemble des libellés à coder. Jusqu'à présent, la codification étant assurée par Sicore qui n'attribue pas systématiquement un code à chaque observation, une phase de reprise manuelle était obligatoire pour coder l'ensemble des bulletins d'une enquête annuelle du recensement. Au contraire de Sicore, les classifieurs que nous avons entraînés grâce à des méthodes d'apprentissage supervisé renvoient en sortie une probabilité d'appartenance à chaque code de la nomenclature. Le code sélectionné pour chaque observation est ainsi le code associé à la probabilité la plus élevée. Si on choisit d'introduire une reprise manuelle en aval du classifieur, il est naturel d'y envoyer les libellés de professions pour lesquels le modèle affiche la confiance la plus faible. Dans la suite, on quantifie la confiance du modèle par la différence entre les probabilités estimées d'appartenance à la classe la plus probable et à la seconde classe la plus probable. Comme le classifieur aura tendance à moins bien classer les observations pour lesquelles il affiche une confiance moins élevée, sa précision devrait augmenter sur le sous-ensemble d'observations non-envoyé en reprise.

Pour évaluer la précision de l'ensemble du codage automatique en présence de reprise, il faut pouvoir évaluer la précision de la reprise sur un sous-ensemble d'observations. Les jeux de données de test utilisés pour ce travail ont aussi été annotés en double codage à l'aveugle, avec arbitrage lorsque les deux codes proposés étaient différents. On évalue donc la précision d'une reprise hypothétique en simple codage comme suit : une observation de notre jeu de données envoyée en reprise manuelle a sa profession « codée correctement » lorsque le code donné par le premier annotateur est identique à celui donné par le second annotateur, ou par le troisième annotateur lorsque ce dernier a dû arbitrer. La précision de la codification pour l'ensemble des libellés est ensuite calculée comme une moyenne pondérée de la précision du classifieur et de la précision de la reprise.

Les précisions du codage par le classifieur (avec variables annexes intégrées à un libellé enrichi), du codage en reprise manuelle et du codage total en fonction du pourcentage d'observations envoyé en reprise pour la profession actuelle salariée est donné en Figure 9. La précision globale

	PROFS	PROFI	PROFA
Dimension de l'espace de plongement des libellés	150	100	100
Taille maximale des n-grammes de mots inclus dans le calcul du plongement des libellés	3	3	3
Taille maximale des n-grammes de caractères inclus dans le calcul du plongement des libellés	4	5	5
Taille minimale des n-grammes de caractères inclus dans le calcul du plongement des libellés	3	3	3
Dimension de l'espace de plongement de la situation	3	3	4
Dimension de l'espace de plongement du statut actuel	3	3	-
Dimension de l'espace de plongement du statut antérieur	-	-	3
Dimension de l'espace de plongement de la position professionnelle	3	-	-
Dimension de l'espace de plongement de la tranche de salariés employés	-	3	-
Dimension de l'espace de plongement de l'activité	70	60	-
Dimension de l'espace de plongement de la catégorie juridique	18	12	-
Dimension de l'espace de plongement de la tranche d'effectifs	3	-	-

TABLE 5 – Hyperparamètres choisis pour les classifieurs avec variables annexes considérées à part.

	PROFS	PROFI	PROFA
Variables annexes concaténées au libellé	66,8 %	69,8 %	70,8 %
Variables annexes considérées à part	66,0 %	69,2 %	71,0 %

TABLE 6 – Proportions pondérées d'observations bien classées pour les deux modèles et chaque type de profession.

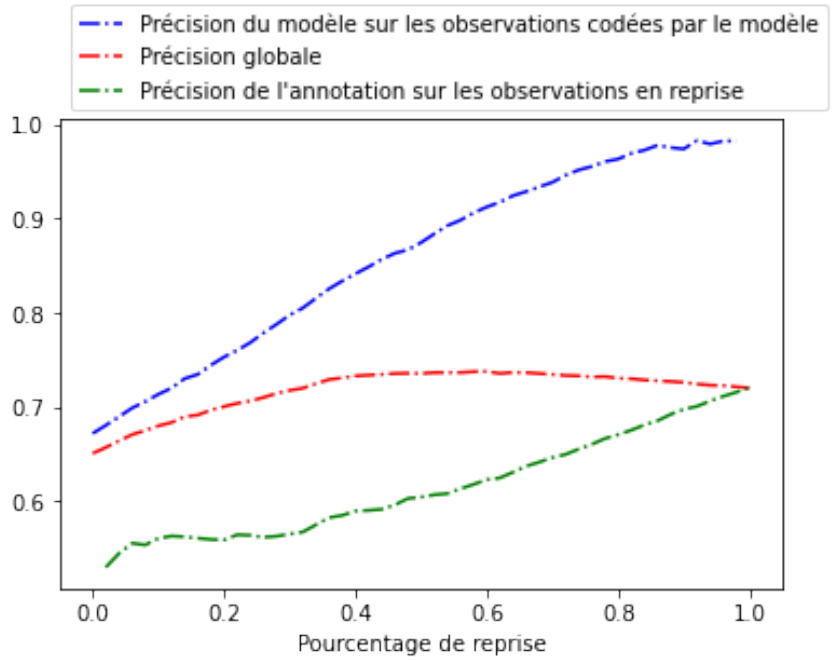


FIGURE 9 – Précision globale de la codification des libellés de profession actuelle salariée.

augmente assez fortement lorsqu'on commence à envoyer des libellés en reprise car le modèle présente un indice de confiance très faible pour ces derniers et renvoie souvent un code erroné. Même si ces libellés sont également difficiles à coder manuellement, pour eux la précision de la reprise manuelle est meilleure que la précision du modèle. Ainsi avec un taux de reprise de 40 %, la précision globale de la codification dépasse 73 %. Le constat est très similaire pour les autres types de profession. Il y a un arbitrage à faire entre les ressources allouées à la reprise manuelle et la précision que cette reprise apporte sur une campagne de codification, au regard des objectifs de qualité globaux.

5 Discussion et suites

Le problème de classification considéré dans ce travail est difficile en grande partie à cause du fort bruit dans les données annotées (65 % des professions de l'étalon-or étaient bien codées par les annotateurs avant la campagne de labellisation), de la relativement faible taille des jeux d'apprentissage à notre disposition et du nombre important de classes de la nomenclature PCS (certaines classes ne sont que très faiblement représentées dans les données).

Une première piste à explorer pour améliorer les résultats affichés en 4 est de complexifier le classifieur utilisé après la phase de plongement, en particulier lorsque les variables annexes

sont intégrées séparément. Dans le cas du classifieur utilisant les libellés enrichis, des relations complexes entre variables sont déjà capturées grâce à la prise en compte des n-grammes de mots. Plusieurs options ont été testées, dont l'utilisation d'un perceptron multi-couches à la place du classifieur linéaire, sans succès. Un travail plus poussé doit être fait sur l'optimisation des hyperparamètres, toujours pour le modèle où on considère les variables annexes à part. Pour pallier la faible taille des jeux d'apprentissage, la reprise manuelle pourra être utilisée au fur et à mesure des enquêtes du recensement pour enrichir les données disponibles et d'autres sources de données pourront être mobilisées. Ces options sont détaillées en 5.2. Enfin, il semble pertinent de mobiliser la structure hiérarchique de la nomenclature PCS qui n'est pas exploitée dans la méthodologie exposée en 2.

5.1 Prise en compte de la hiérarchie de la nomenclature pour améliorer la précision

Les classifieurs décrits précédemment sont construits de telle sorte que la couche en sortie représente le vecteurs de probabilités « one-vs-all » de chacun des codes PCS possibles sur 1 à 4 niveaux de la nomenclature ou la sortie non-codable. Toutes les issues sont donc présentes « à plat ». La dimension hiérachique de la nomenclature n'est pas prise en compte dans la méthodologie. Plusieurs stratégies existent et ont été explorées afin de tirer profit de cette caractéristique et ainsi améliorer la précision des modèles [13].

Une approche possible (utilisée dans [14]) consiste à construire un modèle par niveau de la nomenclature PCS (le premier modèle cherche à prédire le code PCS au premier niveau de la nomenclature, le deuxième au deuxième niveau etc.). Imaginons qu'on hésite au 4ème niveau entre 2 codes différents, qui diffèrent même au premier niveau de la nomenclature. Dans ces cas-là il est possible que classifieur au premier niveau nous aide à départager. On peut choisir le code PCS qui maximise la combinaison linéaire $\alpha_1 p_1 + \alpha_2 p_2 + \alpha_3 p_3 + \alpha_4 p_4$ avec des coefficients α_i fixés qu'on peut optimiser par *grid search* et p_i la probabilité du code appris par le modèle de $i^{\text{è}}$ niveau. Une autre quantité possible à maximiser est le produit $p_1 p_2 p_3 p_4$. Ces solutions ont donné une précision identique à celle atteinte avec une vision plate de la nomenclature mais n'ont pas permis de la dépasser.

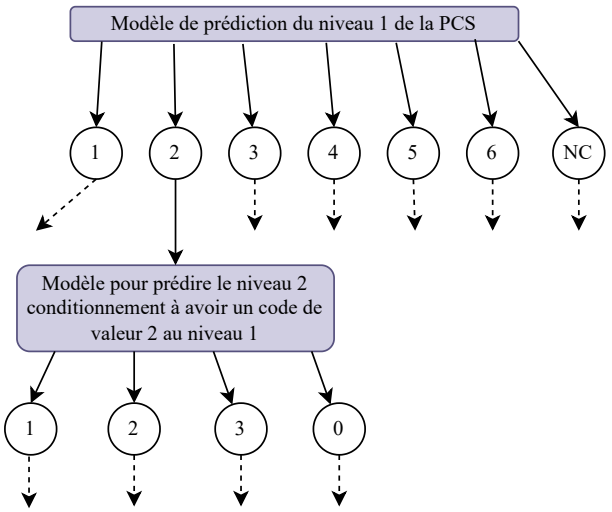


FIGURE 10 – Représentation de la nomenclature sous forme d'arbre de modèles.

Une autre solution est de réaliser la prédiction à partir d'un arbre de modèles. Pour coder, on

descend l'arbre de la nomenclature en ne gardant que les top n branches à chaque niveau. Arrivé aux feuilles de l'arbre, on retient le code avec le produit $p_1p_2p_3p_4$ le plus élevé. Cette stratégie donne une précision inférieure car les modèles plus profonds sont appris sur peu d'observations et la quantité $p_1p_2p_3p_4$ privilégie les branches peu ramifiées.

Des pistes existent pour aller plus loin dans cette direction. Par exemple, [15] propose une approche avec un classifieur pour chaque nœud de la nomenclature considérée qui traite le problème de la taille des jeux d'entraînement pour certains des modèles les plus profonds. [16] donne une méthode adaptée à des problèmes de classification avec un nombre de classes élevé pour lequel une structure hiérarchique existe. Enfin, [17] et [18] présentent des méthodes de classification multi-label hiérarchique, où une observation peut être associée à plusieurs classes de la nomenclature se trouvant à son niveau le plus bas ou non (dans le second cas, qui correspond à celui de ce travail, on parle de *non-mandatory leaf prediction*). Le contexte est différent ici car une observation est associée à une unique classe mais ces méthodes devraient pouvoir être adaptées à notre cas d'étude.

5.2 Vers un passage en production

Ces performances de codification élevées, de 66,8 % à 70,8 % de précision selon le type de profession considéré, ne sont pas directement comparables avec l'existant. En effet, dans le cadre du processus cible imaginé, ces modèles sont exploités sur les libellés les plus difficiles à traiter qui n'ont pu être codés au travers de l'index numérique des professions introduit avec la rénovation de la nomenclature. C'est sur de tels libellés que les classifieurs ont été testés. En combinant les modèles à l'index qui code parfaitement une liste de plus de 5 000 libellés distincts de professions, la fiabilité de la codification de l'ensemble des professions d'une EAR atteindrait plus de 80 % pour les trois types de professions. Ces précisions, estimées pour le processus cible sans même insérer une part de reprise manuelle, sont supérieures à celles mesurées sur le processus actuel en PCS 2003 dans les opérations qualité du recensement. Ces résultats satisfaisants ont conduit à organiser le passage en production de cette méthode.

Cependant, même si la reprise manuelle ne semble pas indispensable pour parvenir à une qualité suffisante pour la codification de l'enquête annuelle de recensement en cours, la labellisation manuelle s'avère essentielle pour acquérir de nouveaux exemples annotés et ainsi améliorer les modèles d'année en année. Une première expérimentation à partir des données des EAR 2015 à 2019 en PCS 2003 semblait indiquer que si aucun ré-entraînement n'avait lieu par absence de reprise manuelle, la précision diminuerait de 2 à 3 points en 4 ans. À l'inverse, en conservant le volume de reprise existant (environ 300 000 bulletins individuels par an), le gain de précision en 4 ans était évalué entre 2 et 4 points. Cette détérioration de la précision au fil du temps sans ajouter de nouvelles données impose une vigilance particulière pour la première utilisation de ces modèles en production à l'horizon 2024, soit 4 ans après la collecte de l'EAR qui a permis d'entraîner la première version des modèles. D'autres sources comme la nouvelle enquête Emploi ou des enquêtes qui embarquent le Tronc Commun des enquêtes Ménages pourraient être exploitées pour enrichir la base d'apprentissage de nouvelles données d'ici la mise en production mais la quantité de données disponibles pourrait ne pas être suffisante pour empêcher une baisse de la fiabilité.

De plus, la reprise manuelle jouera un rôle important pour constituer des échantillons d'évaluation et monitorer les modèles. Un travail important consiste à concevoir la reprise manuelle pour la cible. Il faudra déterminer la répartition entre la part de la reprise consacrée à améliorer la codification de l'année en cours et ré-entraîner les modèles et l'autre part pour constituer un jeu de test et évaluer la précision. Le protocole de reprise pourrait éventuellement être reconsidéré en

choisissant une double codification avec arbitrage pour avoir une annotation de meilleure qualité par rapport à un simple codage sur plus de professions. Concernant le choix des bulletins individuels envoyés en reprise, il sera doublement utile de sélectionner les observations pour lesquelles l'indice de confiance du modèle est bas, car on s'attend à ce que ce soit elles qui améliorent le plus à la fois la qualité de la campagne en cours et le modèle après ré-entraînement.

Références

[1] T. Amossé, O. Chardon, and A. Eidelman, "La rénovation de la nomenclature socioprofessionnelle (2018-2019) : rapport du groupe de travail du Cnis," research report, Conseil national de l'information statistique (Cnis), Dec. 2019.

[2] A. Ikudo, J. Lane, J. Staudt, and B. Weinberg, "Occupational classifications : A machine learning approach," Working Paper 24951, National Bureau of Economic Research, August 2018.

[3] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," *PLOS ONE*, vol. 15, pp. 1–22, 05 2020.

[4] J. Ramos, "Using tf-idf to determine word relevance in document queries."

[5] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, p. 1137–1155, mar 2003.

[6] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *CoRR*, vol. abs/1310.4546, 2013.

[7] J. Pennington, R. Socher, and C. D. Manning, "Glove : Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[8] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *CoRR*, vol. abs/1607.01759, 2016.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT : Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[10] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot, "Camembert : a tasty french language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[11] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *CoRR*, vol. abs/1607.04606, 2016.

[12] B. Settles, "Active learning literature survey," Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[13] C. Silla and A. Freitas, "A survey of hierarchical classification across different application domains," *Data Mining and Knowledge Discovery*, vol. 22, pp. 31–72, 01 2011.

[14] G. Hyukjun, S. Matthias, S. Stefan, K. Lars, and B. Michael, "Three methods for occupation coding based on statistical learning," *Journal of Official Statistics*, vol. 33, no. 1, pp. 101–122, 2017.

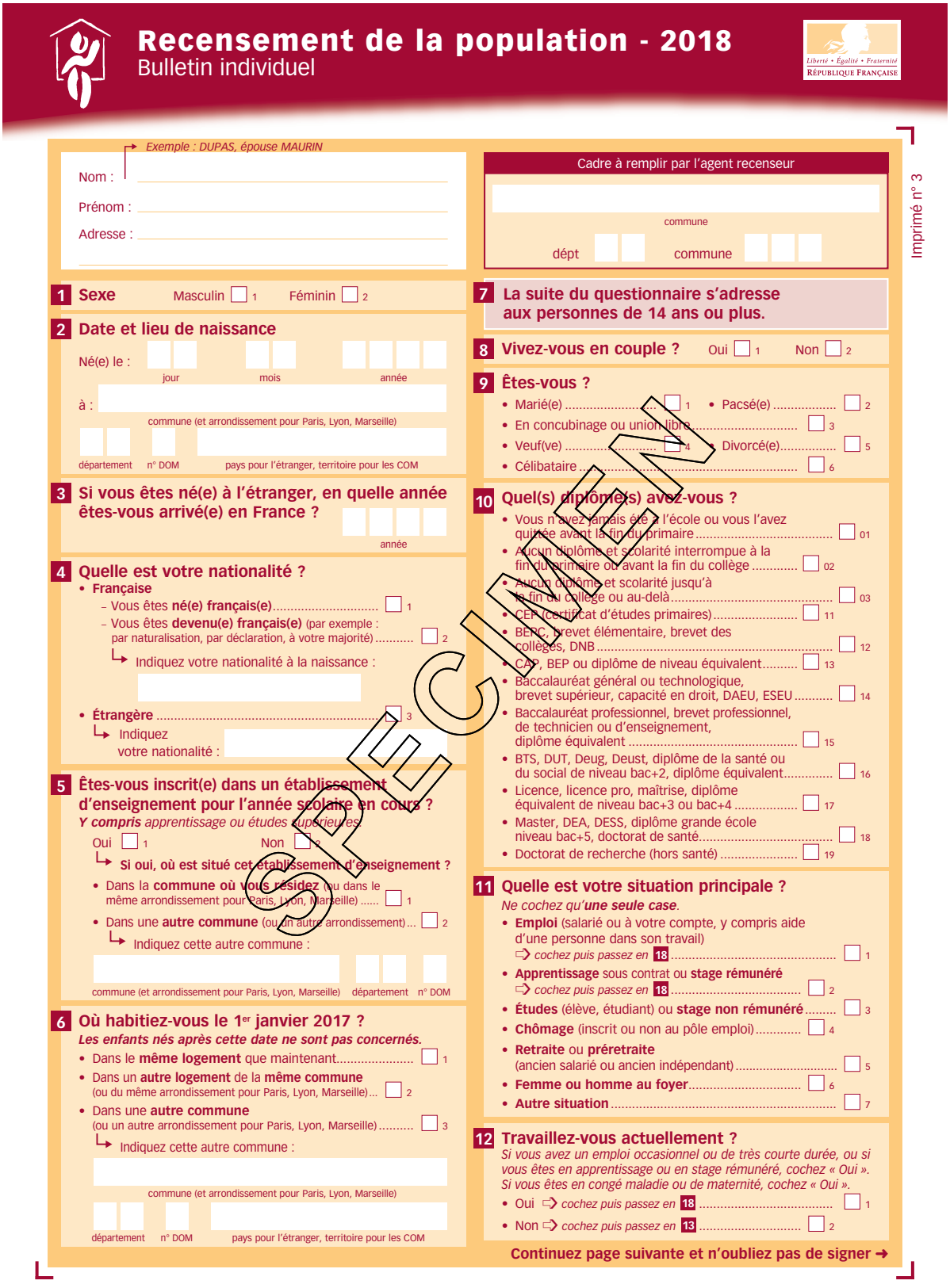
[15] C. Xu and X. Geng, "Hierarchical classification based on label distribution learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 5533–5540, 07 2019.

[16] S. Bengio, J. Weston, and D. Grangier, "Label embedding trees for large multi-class tasks," in *Advances in Neural Information Processing Systems* (J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, eds.), vol. 23, Curran Associates, Inc., 2010.

[17] J. Wehrmann, R. Barros, S. Dôres, and R. Cerri, "Hierarchical multi-label classification with chained neural networks," pp. 790–795, 04 2017.

[18] R. Cerri, R. C. Barros, and A. C. de Carvalho, "Hierarchical multi-label classification using local neural networks," *Journal of Computer and System Sciences*, vol. 80, no. 1, pp. 39–56, 2014.

Annexes



Recensement de la population - 2018
Bulletin individuel

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Imprimé n° 3

Exemple : DUPAS, épouse MAURIN

Nom :
Prénom :
Adresse :

Cadre à remplir par l'agent recenseur
commune
dépt commune

1 Sexe Masculin 1 Féminin 2

2 Date et lieu de naissance
Né(e) le : jour mois année
à : commune (et arrondissement pour Paris, Lyon, Marseille)
département n° DOM pays pour l'étranger, territoire pour les COM

3 Si vous êtes né(e) à l'étranger, en quelle année êtes-vous arrivé(e) en France ? année

4 Quelle est votre nationalité ?
- **Française**
  - Vous êtes **né(e) français(e)** 1
  - Vous êtes **devenu(e) français(e)** (par exemple : par naturalisation, par déclaration, à votre majorité) 2
  - Indiquez votre nationalité à la naissance :
- **Étrangère** 3
  - Indiquez votre nationalité :

5 Êtes-vous inscrit(e) dans un établissement d'enseignement pour l'année scolaire en cours ?
***Y compris** apprentissage ou études supérieures.*
Oui 1 Non 2
- **Si oui, où est situé cet établissement d'enseignement ?**
- Dans la **commune où vous résidez** (ou dans le même arrondissement pour Paris, Lyon, Marseille) 1
- Dans une **autre commune** (ou un autre arrondissement) 2
  - Indiquez cette autre commune :

commune (et arrondissement pour Paris, Lyon, Marseille) département n° DOM

6 Où habitiez-vous le 1er janvier 2017 ?
Les enfants nés après cette date ne sont pas concernés.
- Dans le **même logement** que maintenant 1
- Dans un **autre logement** de la **même commune** (ou du même arrondissement pour Paris, Lyon, Marseille) 2
- Dans une **autre commune** (ou un autre arrondissement pour Paris, Lyon, Marseille) 3
  - Indiquez cette autre commune :

commune (et arrondissement pour Paris, Lyon, Marseille)
département n° DOM pays pour l'étranger, territoire pour les COM

7 La suite du questionnaire s'adresse aux personnes de 14 ans ou plus.

8 Vivez-vous en couple ? Oui 1 Non 2

9 Êtes-vous ?
- Marié(e) 1
- Pacsé(e) 2
- En concubinage ou union libre 3
- Veuf(ve) 4
- Divorcé(e) 5
- Célibataire 6

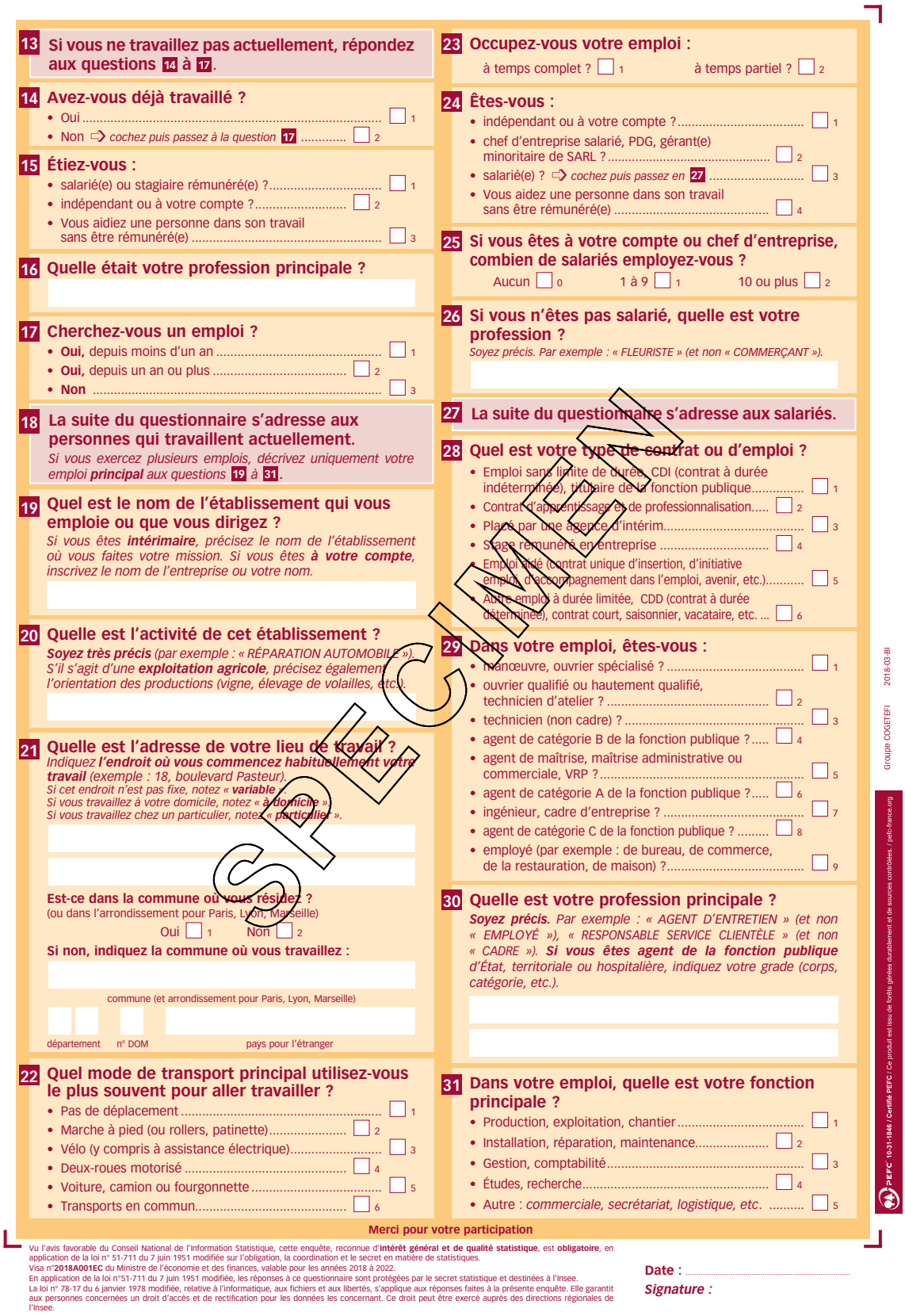
10 Quel(s) diplôme(s) avez-vous ?
- Vous n'avez jamais été à l'école ou vous l'avez quittée avant la fin du primaire 01
- Aucun diplôme et scolarité interrompue à la fin du primaire ou avant la fin du collège 02
- Aucun diplôme et scolarité jusqu'à la fin du collège ou au-delà 03
- CEP (certificat d'études primaires) 11
- BEPC, brevet élémentaire, brevet des collèges, DNB 12
- CAP, BEP ou diplôme de niveau équivalent 13
- Baccalauréat général ou technologique, brevet supérieur, capacité en droit, DAEU, ESEU 14
- Baccalauréat professionnel, brevet professionnel, de technicien ou d'enseignement, diplôme équivalent 15
- BTS, DUT, Deug, Deust, diplôme de la santé ou du social de niveau bac+2, diplôme équivalent 16
- Licence, licence pro, maîtrise, diplôme équivalent de niveau bac+3 ou bac+4 17
- Master, DEA, DESS, diplôme grande école niveau bac+5, doctorat de santé 18
- Doctorat de recherche (hors santé) 19

11 Quelle est votre situation principale ?
*Ne cochez qu'**une seule case**.*
- **Emploi** (salarié ou à votre compte, y compris aide d'une personne dans son travail)
  ⇨ *cochez puis passez en* **18** 1
- **Apprentissage** sous contrat ou **stage rémunéré**
  ⇨ *cochez puis passez en* **18** 2
- **Études** (élève, étudiant) ou **stage non rémunéré** 3
- **Chômage** (inscrit ou non au pôle emploi) 4
- **Retraite** ou **préretraite** (ancien salarié ou ancien indépendant) 5
- **Femme ou homme au foyer** 6
- **Autre situation** 7

12 Travaillez-vous actuellement ?
Si vous avez un emploi occasionnel ou de très courte durée, ou si vous êtes en apprentissage ou en stage rémunéré, cochez « Oui ». Si vous êtes en congé maladie ou de maternité, cochez « Oui ».
- Oui ⇨ *cochez puis passez en* **18** 1
- Non ⇨ *cochez puis passez en* **13** 2

Continuez page suivante et n'oubliez pas de signer ➜

FIGURE 11 – Page 1 du bulletin individuel pour le recensement de la population en 2018.



13 Si vous ne travaillez pas actuellement, répondez aux questions **14** à **17**.

14 Avez-vous déjà travaillé ?
- Oui ☐ 1
- Non ⇨ *cochez puis passez à la question* **17** ☐ 2

15 Étiez-vous :
- salarié(e) ou stagiaire rémunéré(e) ? ☐ 1
- indépendant ou à votre compte ? ☐ 2
- Vous aidiez une personne dans son travail sans être rémunéré(e) ☐ 3

16 Quelle était votre profession principale ?

17 Cherchez-vous un emploi ?
- **Oui,** depuis moins d'un an ☐ 1
- **Oui,** depuis un an ou plus ☐ 2
- **Non** ☐ 3

18 La suite du questionnaire s'adresse aux personnes qui travaillent actuellement.
*Si vous exercez plusieurs emplois, décrivez uniquement votre emploi **principal** aux questions **19** à **31**.*

19 Quel est le nom de l'établissement qui vous emploie ou que vous dirigez ?
*Si vous êtes **intérimaire**, précisez le nom de l'établissement où vous faites votre mission. Si vous êtes **à votre compte**, inscrivez le nom de l'entreprise ou votre nom.*

20 Quelle est l'activité de cet établissement ?
***Soyez très précis** (par exemple : « RÉPARATION AUTOMOBILE »).*
*S'il s'agit d'une **exploitation agricole**, précisez également l'orientation des productions (vigne, élevage de volailles, etc.).*

21 Quelle est l'adresse de votre lieu de travail ?
*Indiquez **l'endroit où vous commencez habituellement votre travail** (exemple : 18, boulevard Pasteur).*
*Si cet endroit n'est pas fixe, notez « **variable** ».*
*Si vous travaillez à votre domicile, notez « **à domicile** ».*
*Si vous travaillez chez un particulier, notez « **particulier** ».*

Est-ce dans la commune où vous résidez ?
(ou dans l'arrondissement pour Paris, Lyon, Marseille)
Oui ☐ 1 Non ☐ 2
Si non, indiquez la commune où vous travaillez :
commune (et arrondissement pour Paris, Lyon, Marseille)
département n° DOM pays pour l'étranger

22 Quel mode de transport principal utilisez-vous le plus souvent pour aller travailler ?
- Pas de déplacement ☐ 1
- Marche à pied (ou rollers, patinette) ☐ 2
- Vélo (y compris à assistance électrique) ☐ 3
- Deux-roues motorisé ☐ 4
- Voiture, camion ou fourgonnette ☐ 5
- Transports en commun ☐ 6

23 Occupez-vous votre emploi :
à temps complet ? ☐ 1 à temps partiel ? ☐ 2

24 Êtes-vous :
- indépendant ou à votre compte ? ☐ 1
- chef d'entreprise salarié, PDG, gérant(e) minoritaire de SARL ? ☐ 2
- salarié(e) ? ⇨ *cochez puis passez en* **27** ☐ 3
- Vous aidez une personne dans son travail sans être rémunéré(e) ☐ 4

25 Si vous êtes à votre compte ou chef d'entreprise, combien de salariés employez-vous ?
Aucun ☐ 0 1 à 9 ☐ 1 10 ou plus ☐ 2

26 Si vous n'êtes pas salarié, quelle est votre profession ?
Soyez précis. Par exemple : « FLEURISTE » (et non « COMMERÇANT »).

27 La suite du questionnaire s'adresse aux salariés.

28 Quel est votre type de contrat ou d'emploi ?
- Emploi sans limite de durée, CDI (contrat à durée indéterminée), titulaire de la fonction publique ☐ 1
- Contrat d'apprentissage ou de professionnalisation ☐ 2
- Placé par une agence d'intérim ☐ 3
- Stage rémunéré en entreprise ☐ 4
- Emploi aidé (contrat unique d'insertion, d'initiative emploi, d'accompagnement dans l'emploi, avenir, etc.) ☐ 5
- Autre emploi à durée limitée, CDD (contrat à durée déterminée), contrat court, saisonnier, vacataire, etc. ☐ 6

29 Dans votre emploi, êtes-vous :
- manœuvre, ouvrier spécialisé ? ☐ 1
- ouvrier qualifié ou hautement qualifié, technicien d'atelier ? ☐ 2
- technicien (non cadre) ? ☐ 3
- agent de catégorie B de la fonction publique ? ☐ 4
- agent de maîtrise, maîtrise administrative ou commerciale, VRP ? ☐ 5
- agent de catégorie A de la fonction publique ? ☐ 6
- ingénieur, cadre d'entreprise ? ☐ 7
- agent de catégorie C de la fonction publique ? ☐ 8
- employé (par exemple : de bureau, de commerce, de la restauration, de maison) ? ☐ 9

30 Quelle est votre profession principale ?
***Soyez précis.** Par exemple : « AGENT D'ENTRETIEN » (et non « EMPLOYÉ »), « RESPONSABLE SERVICE CLIENTÈLE » (et non « CADRE »). **Si vous êtes agent de la fonction publique** d'État, territoriale ou hospitalière, indiquez votre grade (corps, catégorie, etc.).*

31 Dans votre emploi, quelle est votre fonction principale ?
- Production, exploitation, chantier ☐ 1
- Installation, réparation, maintenance ☐ 2
- Gestion, comptabilité ☐ 3
- Études, recherche ☐ 4
- Autre : *commerciale, secrétariat, logistique, etc.* ☐ 5

Merci pour votre participation

Vu l'avis favorable du Conseil National de l'Information Statistique, cette enquête, reconnue d'**intérêt général et de qualité statistique**, est **obligatoire**, en application de la loi n° 51-711 du 7 juin 1951 modifiée sur l'obligation, la coordination et le secret en matière de statistiques.
Visa n°**2018A001EC** du Ministre de l'économie et des finances, valable pour les années 2018 à 2022.
En application de la loi n°51-711 du 7 juin 1951 modifiée, les réponses à ce questionnaire sont protégées par le secret statistique et destinées à l'Insee.
La loi n° 78-17 du 6 janvier 1978 modifiée, relative à l'informatique, aux fichiers et aux libertés, s'applique aux réponses faites à la présente enquête. Elle garantit aux personnes concernées un droit d'accès et de rectification pour les données les concernant. Ce droit peut être exercé auprès des directions régionales de l'Insee.

Date : .....
Signature :

Groupe CGETEFI 2018-03-BI

PEFC 10-31-1846 / Certifié PEFC / Ce produit est issu de forêts gérées durablement et de sources contrôlées. / pefc-france.org

FIGURE 12 – Page 2 du bulletin individuel pour le recensement de la population en 2018.