OTTAWA 2023
64TH WORLD STATISTICS CONGRESS

Machine learning for coding occupations in the Census: first lessons from experiments to production[1]

Théo Leroy, Institut National de la statistique et des études économiques (INSEE), theo.leroy@insee.fr

Lucas Malherbe, Institut National de la statistique et des études économiques (INSEE), lucas.malherbe@insee.fr

Tom Seimandi, Institut National de la statistique et des études économiques (INSEE), tom.seimandi@insee.fr

Elise Coudin, Institut National de la statistique et des études économiques (INSEE) and CREST, elise.coudin@insee.fr

Abstract

This paper presents the approach undertaken by INSEE to select and implement classification of the occupational variables of the annual census survey in the new national occupational classification (PCS 2020). The coding process will use a combination of automatic approaches (list auto-completion and supervised ML prediction models) and manual coding. An ad hoc annotation campaign conducted in 2021 provides a first set of training and testing of the algorithms. A two-layer neural network algorithm (fastText embeddings of words and n-grams and classifier) allows to achieve overall accuracy goals fixed as conditions for going into production.

Keywords: Experimentation, occupational classification, NLP, supervised machine learning, automatic coding, fastText.

[1] For helpful comments and discussions, we thank Hélène Chaput, Tristan Loisel, Amandine Schreiber, Patrick Sillard, Gwennael Solard, census and PCS teams. Previous versions of this paper were presented at the New techniques and technologies for official statistics (NTTS) 2021 conference, at the Journées de la Méthodologie Statistique 2022. This paper reflects the authors' opinions and does not express those of their institutions.

1. Introduction

Occupational classifications are useful to provide statistic, economic and social descriptors both accounting for similarities in job tasks and contents and similarities in economic and institutional contexts. To provide realistic social and economic analyses, occupational classification dictionaries have to be regularly updated. In 2020, a new dictionary of the French occupation classification (PCS 2020) was disseminated, accompanied with an auto-completion tool, which links perfectly a list of the most frequent jobs to their classification class. It is planned to use the PCS 2020 dictionary since 2024 in the annual census survey. However, responses not in this list remain to be classified. INSEE has chosen not to adapt its rule-based automatic coding system set to classify within the previous dictionary (PCS 2003) to the new dictionary. INSEE rather has chosen to experiment the use of machine learning techniques to perform this text classification task for which they are expected to perform well. In 2021, a large campaign of manual labeling was conducted, with the aim of ensuring the quality of the training and test sets on which the algorithms would be trained or tested. A two-layer neural network algorithm (fastText embeddings of words and n-grams and classifier) was trained and finally selected. The combination of the two automatic coding modes (list and supervised learning on non-lists) seems to exceed the accuracy rates of the previous system at the finest level for the current occupation. However, performances are lower for classifying the past occupation declared by retirees and unemployed people, which counts more paper questionnaires. The combination with a part sent to be re-coded by human annotators allows to gain some points of accuracy. Based on these results, the integration of predicting and training tools into the census production chain is investigated in 2022. This paper retraces this experience. Section 1 reports some contextual elements. Section 2 presents more precisely the data and the methods used. Results and evaluation are exposed in section 3. Section 4 is dedicated to questions relative production integration. Section 5 concludes.

2. Context

2.1. Renovation of the French occupational classification

Since its creation in the early 50s, the "Professions and catégories socio-professionnelles" (PCS) classification has been updated every 20 years to account for structural changes of the occupational landscape. New dictionaries were regularly elaborated in 1982, 2003, and, the last one in 2020. In 2018-2019, a working group from the National Council for Statistical Information led the most recent renovation, following four main objectives, see Amossé, Chardon, and Eidelman (2019):

- renewing coding procedures within the PCS classification in order to simplify and standardize production and collection of data in all surveys, as well as, in the long term, within other socio-occupational classifications such as ISCO or ESeG;

- updating the dictionary at the finest 4-digit level in order to account for the evolution of occupations, and providing a detailed up-to-date labor market analysis grid. The first digit and the 2-digit aggregated levels remain unchanged to insure comparability over time;

- defining, for individuals such as households, additional groups of occupations to complete the possibility of analysis. The renovation introduced an 3-digit level, which can be interpreted, together with four transversal groups of occupations of particular interest: the teaching occupation group, the group of occupations related to digital, green jobs, and the group of executives, professionals and high-level experts;

- improving the documentation and dissemination of information relating to the classification in order to facilitate its appropriation by a wide range of users.

The new dictionary remains organized with a tree logic starting from 6 very large groups, subdivided into 31 socio-professional categories (2 digits), themselves subdivided into 126 3-digit groups, and into 316 occupation classes (4 digits). Compared to the 2003 dictionary, there is therefore the creation of this 3-digit grouping level and a major overhaul of the 4-digit level (occupations). Occupation classes are less numerous (316 against 486 previously), revised, and of more homogeneous sizes.

In parallel with this renovation, the working group proposed a list of enriched and standardized job descriptions, which, when combined to only three auxiliary variables - status, position and firm size - allows for unambiguous coding in a 4-digit class. The introduction of additional information directly into the job description simplifies the coding process. In particular, job descriptions are enriched with industry when needed for classification. Hence, a "acheteur approvisionneur de l'industrie", is differentiated from a "acheteur approvisionneur du commerce", which alleviates the need to rely on the full economic activity classification.

An auto-completion tool has also been developed to easily search within this list of around 5800 enriched job descriptions. If no enriched job description is suitable, the respondents are invited to enter their job descriptions in clear text. These clear texts are then sent to human coding. This tool has been used since 2021 for the labor force survey collection.

The approach is particularly suitable for online and computer-assisted surveys for which the auto-completion tool is available and the number of responses to be coded manually (those not in the list) naturally remains low. However, it raises challenges for voluminous operations or operations relying heavily on paper; like the population census.

2.2. Coding occupations in the annual census survey

In the annual census survey (2019), around 2,9 millions of people declare a present or past occupation, around 60% through web questionnaires, 40% through paper questionnaires. Only in 35% of the paper questionnaires, job textual information match one job description in the list of enriched job descriptions, which allows a direct and unambiguous coding in the 2020 PCS dictionary. The rate of use of the list for web questionnaires is not known and depends heavily on the ergonomics of the collection medium (smartphones, computers, tablets). In the labor force survey, the first survey that uses the PCS 2020 auto-completion tool, 80% of the respondents choose a label in the list whereas 20% declare her/his occupation in clear text. Combining those figures leads to having a direct coding thanks to the list only for 60% of the census responses, whereas the actual PCS 2003 automatic coding system handles 88% of cases. An automatic coding tool is then needed to handle the remaining, as the manual coding cannot exceed 12% of cases (resource constraints).

The current PCS 2003 automatic coding process relies on an meta-expert system with deterministic rules (système informatique de codage aux enquêtes, SICORE) parameterized for coding into the PCS 2003 dictionary. SICORE is used for various classification tasks at INSEE, see Rivière (1995), Schuhl (1996). After a first step of pre-processing of the textual job description, a normalized job description is searched into an index of around 100,000 jobs, each of them associated with one or several potential PCS 2003 codes (called pre-codes). The determination of those pre-codes is done by going through an optimized decision tree based on the n-grams composing the job description, see Lorigny (1988). Then, to decide between the potential codes, a set of deterministic rules involving auxiliary variables is used. The PCS 2003 coding relies on 10 auxiliary variables. The case distinction is done through another decision tree of more than 17,000 nodes, see Leroy (2022). When SICORE does not find a match, the observation is sent for manual re-coding. In practice, each year, 12% of the annual census questionnaires are sent to manual re-coding, the automatic system being not able to code them. SICORE is an expert system, *i.e.* the reference indexes, the logical rules involving auxiliary variables are specified and maintained by classification experts.[2] The complexity, accumulation, and interweaving of decisions makes the system difficult to modify in practice to account for profound changes in classifications. This situation gave us the opportunity to experiment the performance of supervised machine learning alternatives.

[2] To be precise, the initialization of the reference index relied on train samples in a very similar way to a machine learning approach, but then additions and modifications of the reference index were done by hand by classification experts.

2.3. Reproducing the 2003 PCS automatic coding system rules with machine learning

A first experiment took place in 2020, see Leroy and Loisel (2021). It consisted in testing the performance of various supervised machine learning algorithms to classify into the PCS 2003 dictionary, and to estimate the minimal size of the data set needed to be labeled in PCS 2020 for initiating an algorithm with an accuracy (% of well-predicted observations) above 80%. The accuracy of 80% corresponds to an overall quality estimate of the actual PCS 2003 process (automatic coding + manual re-coding), which is regularly evaluated thanks to quality control surveys during which a sub-sample of questionnaires are re-coded twice, and a third time when the first two codes differ.

Linear SVM with TF-IDF embeddings, random forests, naive bayes, k-neighbors and a two-layer neural network (with word and sub-word embeddings+ classifier) based on the fastText library were compared. Only the latter shows an accuracy above 80% for a training set of 79,000 questionnaires, and its accuracy stays higher than other ones even with larger training set sizes.

The minimum sizes of the train and test samples to meet the quality constraints, in terms of automatic coding rate and accuracy were estimated at 85,000 questionnaires. An initial set composed of the list of enriched job descriptions plus around 100,000 questionnaires labeled in PCS 2020 would enable one to train and test the algorithm to code in the new dictionary, providing an accuracy higher than 80%. The experiment showed also that the training set does not need to be representative of the distribution of jobs in the population, but rather to cover the largest range of job descriptions and combinations of auxiliary variables as possible.

Last teaching, when the model is trained with data of a given year, say y, its prediction accuracy decreases for subsequent years: at $y + 4$, it looses 4 points of accuracy. Regular re-coding and model re-training are needed to maintain the model up-to-date, which will be a challenge for production integration. However, at this stage, the potential of a machine learning approach had been demonstrated and it was convincing enough to unlock resources for a first *ad-hoc* coding campaign. Well, in practice we took advantage of an opportunity linked to the COVID/lockdown period.

2.4. An ad hoc labeling campaign

Due to the COVID-2019 crisis and periods of lockdown in France, the 2021 census survey, which should have taken place in January 2021, was postponed to 2022. Instead, the census teams are available to participate to various other operations, among which the large one-shot labeling campaign in PCS 2020 mentioned above. This campaign took place during first semester of 2021. Around 120,000 census job answers were classified in PCS 2020, each twice, by two different human annotators, and a third one for trade-off when required. This dataset constitutes train and test sets for supervised machine learning algorithms. The double coding + trade-off aim to ensure the quality of the training and test sets on which the algorithms would be trained or tested. We come back on the approach followed to construct the train and test samples in section 4.

3. Data and methods

This section provides deeper details on the data, the classification methods used and the approach followed for constructing the train/test samples that were labeled during the *ad hoc* coding campaign.

3.1. Data

There are three types of job descriptions to be coded in occupations, collected in the annual census survey, depending on the employment situation of the individual:

- current occupation for wage-earners (PROFS), based mainly on the textual response to the question "*What is your main occupation?*" "*Quelle est votre profession principale ?*"

- current occupation for self-employed (PROFI), based mainly on the textual response to the question "*If you are not wage-earner/employee, what is your occupation?*" "*Si vous n'êtes pas salarié, quelle est votre profession?*"

- past occupation for retirees and non-employed, based on the textual response to the question "*What was your main occupation ?*" "*Quelle était votre profession principale ?*"

Six auxiliary variables are needed for classifying wage-earner occupations into the PCS 2020 classification, five for self-employed, and two for past occupations, see Table 1, and the questionnaires in the appendix. These variables are categorical and mainly correspond to additional information on position/qualification, economic activity, form of economic and financial control, size of the firm or local unit. The economic activity and the form of economic and financial control used are classes in the corresponding classifications, and can take many different values (177 economic and financial control forms and 718 classes of economic activity are present in the data).

The choice has been made to have algorithms relying only on this subset of variables - textual job description + auxiliary variables such as defined in Table 1 even though some AI algorithms could in theory retrieve information usable for classification from other parts of the census questionnaire. This choice is guided by both practical and ethical considerations. It reduces the black box effect of AI algorithms, and the risks of uncontrolled bias based on other non-anticipated correlations.

		Occupation type		
Auxiliary variables		Wage-earners	Self-employed	Past
Employment status	• Employed • Apprentice • On going studies • Unemployed • Retiree • Non-participant • Other	✓	✓	✓
Current status	• Self-employed • Entrepreneur, chief executive • Employee • Family worker	✓	✓	
Past status	• Employee or paid-intern • self-employed • family worker			✓
Position	• Laborer • Skilled blue collar, • Technician • Civil-servant of category B • Associate professional • Civil-servant of category A • Professional, Executive, Manager • Civil-servant of category C or D • White-collar worker	✓		
Number of employed workers	• 0 • 1 to 9 • 10 and above		✓	
Economic activity	Codes of the national classification	✓	✓	
Form of economic and financial control	Codes of the national classification	✓	✓	
Number of workers in the local unit		✓		

Table 1: Auxiliary variables used for classifying into the PCS 2020 dictionary.

Finally, the objective is simply to classify the set composed of the textual job description and the auxiliary variables into the PCS 2020 dictionary. If the job description enriched with information from the auxiliary variables corresponds to an element of the list of the enriched job descriptions, the coding is direct and unambiguous. If not, a predictive task is needed. There are three predictive tasks to train, one per type of occupations. We describe below the main principle of the predictive algorithms used. Such as the deterministic part, the predictive tasks only use textual job description and auxiliary variables presented in Table 1.

3.2. Supervised learning with embedding for text classification

The methodological approach followed is detailed in Leroy, Malherbe, and Seimandi (2022), we report below a synthetic summary. The algorithm found to far outperforms other investigated methods in various performance dimensions relies on the open-source fastText library,[3] (Joulin, Grave, Bojanowski, and Mikolov, 2016, Joulin, Grave, Bojanowski, Douze, Jégou, and Mikolov, 2016). Its architecture consists in a two-layer neural network. The first layer consists in representing words and documents in low dimension vectors: document features (words and subwords) are embedded in low dimensional vectors and their representations are averaged for representing the document. Those then feed a linear classifier activated by a function f to compute the probability distribution over the predefined classes. This leads in minimizing the negative cross-entropy over the classes:

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

where x_i is the set of words in the document i, y_i the known class of document i, A and B two weight matrices. A, the embedding matrix (word embeddings + document average), and B, the classifier matrix, are learned simultaneously during training.

The linear classifier B takes the dense vector representation of the input observation $w_i = Ax_i$ to predict a class by applying an activation function f. Practical considerations lead us to opt for a *one-versus-all* strategy, according to which K independent binary classifiers are trained, one for each class (vs all other classes). The activation function is composed of K sigmoid functions

$$f(z) = \left[\frac{\exp(z_1)}{1+\exp(z_1)}, ..., \frac{\exp(z_K)}{1+\exp(z_K)}\right],$$

[3] `https://fasttext.cc`

where K is the number of classes, $z = Bw_i$ with w_i the vector representation of observation i. By systematically choosing the code corresponding to the highest probability, the scheme *one-versus-all* gives better results than the *softmax* function usually used for multinomial choices.

Let us go back to the feature engineering/embedding stage. The main improvement of the method used here compared to a standard *bag of words* model is that the document vector representations are based on both the words that compose the document and a series of both n-grams of words and n-grams of characters (subwords), see Bojanowski, Grave, Joulin, and Mikolov (2016).

For example, RESPONSABLE MAGASIN LOGISTIQUE will be represented by the average of the vectorial representations of its words RESPONSABLE, MAGASIN, LOGISTIQUE, plus bigrams of words RESPONSABLE MAGASIN and MAGASIN LOGISTIQUE, plus trigrams of words RESPONSABLE MAGASIN LOGISTIQUE, plus subwords (n-grams of characters). Here, with the example of 3-grams of characters: <RE, RES, ESP, SPO, PON, NSA, ABL, BLE, LE>, <MA, MAG, AGA, GAS, ASI, SIN, IN>, <LO, LOG, OGI, GIS, IST, STI, TIQ, IQU, QUE, UE>, plus subwords of bigrams of words : LEM, EMA, INL, NLO. The characters $<$ and $>$ stand for the word begining and ending. This n-gram representation is quite robust to spelling or typing errors. This also enables one to compute a vector representation for documents or words not present in the training corpora. This enables one to propose a class for occupations not previously seen, based on a sort of semantic proximity of the n-grams that compose them. Further, additional features such as LEM, EMA, INL, NLO, in the previous example, capture some partial information about the order of the words and the interactions of words in the document, without explicitly modeling them.

Joulin, Grave, Bojanowski, and Mikolov (2016) indicates that the method used here gives results for a document classification task comparable to much more complex models such as BERT, (Devlin, Chang, Lee, and Toutanova, 2019) or the derived model CamemberBERT (Martin, Muller, Ortiz Suárez, Dupont, Romary, de la Clergerie, Seddah, and Sagot, 2020) for French, which have *state-of-the-art* performance for many language processing tasks, while being much faster.

The maximal size of n-grams of words, n-grams of characters, the minimal size of n-grams of characters together with the dimension of the vector representation (embeddings) are hyperparameters optimized during the training of the model. The fastText library is scalable and optimized to train models using few computing resources (it does not require GPU) and adapt to a large volume of data.

In practice, to account both for the occupation wording and the auxiliary variables needed for classification, the occupation wording and all the auxiliary variables are concatenated into a single "enriched occupation wording", which in turn is decomposed into words, n-grams of words and of subwords as described above and embedded in a single vector space. With two auxiliary variables, this would yield for our previous example RESPONSABLE MAGASIN LOGISTIQUE SITUATION_1 POSITION_9. This approach allows to easily reduce the dimension of some sparse auxiliary, such as the industry classified at a fine detail in the classification of economic activities dictionary. However, this strategy has flaws. The concatenation appears in a given order, which can be problematic to account for interactions between variables that do not appear in the n-grams of words because they are too far in the enriched wording. The semantic behind the codes of the auxiliary variables is not accounted for, while it could help especially since they undergo the same embedding or on the contrary could introduce too much noise. The approach imposes a structure of concatenation, which reduces its possibility to fit other survey needs (where response modalities are just slightly different).

Other options could correct those flaws, such as training/using different embeddings for the different variables, later concatenated to feed the classifier ; accounting for the complete semantic, using embeddings coming from outside data sets (for the activity for example)... However, in practice, those alternatives have not shown yet better results than those obtained with the crude concatenation.

Confidence index. The classifiers described above predict the probabilities of belonging to each class of the PCS dictionary for any observation given as input. We retain as prediction the class associated with the highest probability. The higher this probability, the more confident the classifier is about its prediction. We use as confidence index the difference between the predicted probability for the most probable class and the predicted probability for the second most probable class. This confidence index goes between 0 (zero confidence) and 1 (total confidence). It indicates how much the model discriminates between classes. We will see later how it can be used to choose which observations have to be sent for human re-coding either to increase the coding campaign quality or to provide additional labeled information to re-train the model.

4. Results and evaluation

4.1. Sampling of train and test samples

Census teams were able to code manually 120,000 census questionnaires, which exceeds slightly the 100,000 questionnaires recommended by the first experiment. The question then was to breakdown those 120,000 observations among the three types of occupation to predict, in order to homogenize the accuracy of the three predictive models, and among training and test samples. The sampling approach was chosen for maximizing the information within the training set given the constraint of 120,000 questionnaires, and while maintaining robustness in the evaluation of model performance. Several sampling strategies were considered and their accuracy into predicting PCS 2003 codes were compared, see Leroy, Malherbe, and Seimandi (2022):

1. random sampling with inclusion probabilities representing the occurrence frequencies of the combination of job description and the most discriminant variables for predicting the PCS 2003, and supposedly the PCS 2020, see Table 2; then only one observation per combination is retained to avoid duplicates;

2. random sampling with inclusion probabilities based on clustering of the same combination of variables represented by pre-trained embeddings;

3. an active learning approach, during which models are continuously retrained with as additional observations those for which they are the less confident in their predictions.

4. a mix of 1 and 3.

Occupation	Sampling variables
Wage-earners	• Textual job description • Occupational status • Position • First digit of the economic activity
Self-employed	• Textual job description • Occupational status • Employed people or not • Number of employed people • First digit of the economic activity
Past occupation	• Textual job description • Occupational status

Table 2: Most discriminant variables for predicting PCS 2003 used as sampling variables.

The most performing approaches appear to be approach 1 and approach 4. The full active learning strategy did not show better accuracy for these size orders, neither the approach based on pre-trained general embeddings. The approach 4 was chosen. A first sample of size n_1 was selected by systematic sampling with inclusion probabilities equaling the occurrence frequencies of the combination of the most discriminant variables for predicting PCS 2003. Then, a classifier was trained on this sample, and predictions were derived for each observation of the sampling frame (exception made of the first sample and of the test sample). Then, a second sample of size n_2 was selected within the observations whose predictions were the most uncertain in terms of confidence index.

Further, the test sample was chosen to be totally disjoint from the training sample. It was selected by systematic sampling with inclusion probabilities equaling the occurrence frequencies of the combination of the sampling variables such as presented in Table 2. The 120,000 observations were therefore broken down as follows

	Samples		
	Training		Test
	n_1	n_2	
Current occupation for wage-earners	70,000	20,000	5,000
Current occupation for self-employed	10,000	2,500	2,000
Past occupation	6,000	1,500	2,000

Table 3: Sizes of samples to be annotated during the *ad-hoc* campaign

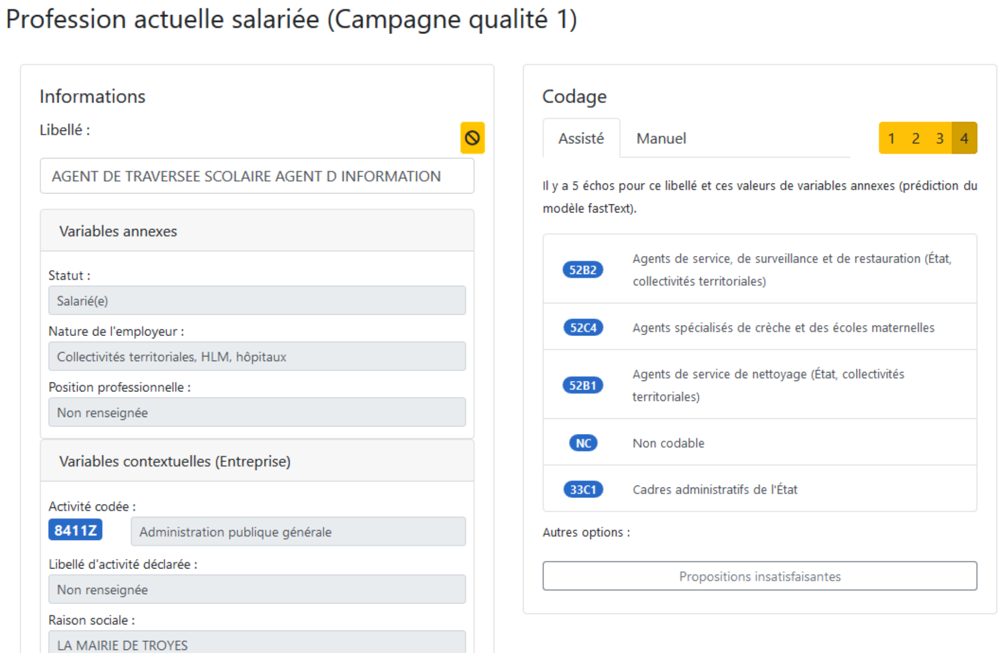
4.2. Ensuring quality of manual labeling during the ad hoc *campaign.*

Supervised classification methods require that the training sample be of very good quality. Hence, occupational classification experts trained annotators to code with the new dictionary. The training session consisted of a day of theory and a half-day of practical cases. Before accessing the coding interface for the 120,000 questionnaires, the annotators were also required to classify a gold standard of 20 cases to ensure that the instructions given during training were well understood. On average, 13 out of 20 occupations were classified correctly at the finest level. This accuracy of 65% is in line with levels typically measured in quality campaigns. Then, training and test data were labeled twice: two annotators each proposed a class blindly. If the classes were identical, the annotation was completed. If not, a third annotator determined the class to be retained, knowing the first two annotations. The class finally chosen could be different from the ones proposed by the first two annotators.

An *ad hoc* application was designed to facilitate annotation, collect and distribute data, ensure that annotators are different for a given questionnaire, and simplify follow-up. This interface presented to the annotator on the left panel the textual job description, the values of the auxiliary variables, and some context variables for difficult cases (like the company name), see Figure 1. The annotation/labeling had to be done on the right panel by selecting a PCS 2020 class. The choice of the class could be done by using

- suggestions from a classifier trained on PCS 2003 data and then converted in PCS 2020 for unambiguous cases; recommended for very simple cases only;

- an autocompletion tool enabling to reconcile if possible the job description to one of the list of enriched job descriptions. This tool could served in case of spelling or typing errors; it was recommended during the training sessions;
- or totally manually, with the possibility to easily see the class description and definition. It was recommended to select a 1- 2- or 3-digit class if the information in the questionnaire were not sufficient to classify at the finest level.



Profession actuelle salariée (Campagne qualité 1)

Informations

Libellé :

AGENT DE TRAVERSEE SCOLAIRE AGENT D INFORMATION

Variables annexes

Statut :
Salarié(e)

Nature de l'employeur :
Collectivités territoriales, HLM, hôpitaux

Position professionnelle :
Non renseignée

Variables contextuelles (Entreprise)

Activité codée :
8411Z Administration publique générale

Libellé d'activité déclarée :
Non renseignée

Raison sociale :
LA MAIRIE DE TROYES

Codage

Il y a 5 échos pour ce libellé et ces valeurs de variables annexes (prédiction du modèle fastText).

52B2	Agents de service, de surveillance et de restauration (État, collectivités territoriales)
52C4	Agents spécialisés de crèche et des écoles maternelles
52B1	Agents de service de nettoyage (État, collectivités territoriales)
NC	Non codable
33C1	Cadres administratifs de l'État

Autres options :

Figure 1: Screenshot of the labeling interface used during the ad-hoc campaign

The campaign involved 55 FTE for 3 months, without counting the time for training the annotators nor for developing the labeling interface.

4.3. Descriptive statistics on train and test sets

There are 316 4-digit classes in the PCS 2020 dictionary, 126 3-digit groups, 31 2-digit groups, and 6 1-digit group. Hence, 480 possible labels if we add a class for unclassified questionnaires, those not informative enough to provide even a 1-digit label. Some groups are specific to wage-earners, some other to self-employed, other mix both status. There are errors also, especially in paper questionnaires, self-employed people can fill the text field relative to salaried occupation or vice-versa: overall, 5% of declared salaried occupations concern self-employed according to the status variable, and 6% of declared self-employed occupations concern wage-earners. The train sample relative to declared salaried occupations contains 428 distinct labels. So, it covers salaried but also some non salaried occupations. The train sample relative to declared self-employed occupations contains 349 distinct labels: more than the number of groups compatible with a non salaried occupation. A first lesson can be learned. The filter between salaried and self-employed occupations based the text filled is not of good quality and part of the annotator job consists of re-allocating declared occupations into their correct category based on the status variables and other information. Any model can only predict a label already seen during training, consequently only errors that were already seen could be corrected. In the future, the questionnaire will be modified and will contain a single text field to declare the current occupation, whether salaried or non salaried. Constructing a single model rather than two different ones according to salaried/non salaried status would be an alternative to seriously consider.

In both train samples, job descriptions are rather short, with 2.4 on average words for declared salaried occupations, and 1.9 for self-employed. Family workers concern 2% of declared self-employed. Some auxiliary variables are missing: the occupational position, the firm economic activity and the firm size, each in 4% of declared salaried occupations, 13% of local unit sizes; and 11% of numbers of employees, 7% of economic activity for declared self-employed occupations. The train sample of declared salaried occupations contains 713 distinct economic activity labels.

Concerning past occupation, which should be coded into 2-digit classes, the train sample contains 35 distinct labels, including the group of observations that cannot be classified, upon the 38 possibilities. Job descriptions contain 2.4 words on average. The past status is missing for 7% of observations.

Quality of train and test data. We report in Table 4 the precision of coding attained by human annotators. In around 72% of cases, a human annotator labels correctly, *i.e.* the code he/she chose corresponds to the one considered as the true one at the end of the coding campaign, the salaried occupation, at the finest level possible,[4] in around 78% for the self-employed, and 77% for past occupations. These figures justify to chose a process with verification - such as two blind codings before trade-off, to approach ground truth. This also shows the inherent uncertainty of these data and of the occupational classification: the classification may be difficult to grasp despite training, the information in the data, not sufficient to discriminate all cases, the final class, subject to a part of interpretation.

	Wage-earners	Self-employed	Past occupation
Annotator 1 - train (unweighted)	71.5 %	76.9 %	76.9 %
Annotator 2 - train (unweighted)	71.6 %	79.7 %	75.9 %
Annotator 1 - test (unweighted)	73.1 %	76.9 %	78.4 %
Annotator 2 - test (unweighted)	73.1 %	77.9 %	78.2 %

Table 4: % of observations well classified by humans

4.4. Models training

Models such as presented in section 2 are trained on the training samples. Hyperparameters are chosen through grid search using a small validating set and reported in Table 5. The training is very rapid, only seconds.

	wage-earners	self-employed	past occupation
Embedding dimension	150	100	100
Maximal size of word n-grams	3	3	3
Maximal size of subword n-grams	4	5	5
Minimal size of subword n-grams	3	3	3

Table 5: Hyperparameters

[4] *i.e* at the 4-digit level if there is enough information, else at the 3-digit level if there is enough information, and so on.

Model accuracy. Table 6 reports model accuracy estimates on train and test data. It appears that the models overfit the train set, this may suggest some overlearning and should be considered more deeply. On test samples, the models tend to achieve between 67% and 70% of correct predictions, which is slightly lower than what one human annotator can do but surely faster.[5]

	Wage-earners	Self-employed	Past occupation
Model accuracy - train (unweighted)	83.2 %	96.0 %	86.9 %
Model accuracy - test (unweighted)	66.1 %	68.0 %	67.1 %
Model accuracy - test (weighted)	66.8 %	69.8 %	70.8 %

Table 6: % of observations well predicted by supervised learning models.

Combination of autocompletion and machine learning prediction accuracy. Next, we estimate the overall performance of the classifying approach, which is: first, classify directly with the autocompletion tool when the job description belongs to the list of enriched job descriptions, and if not possible, use supervised learning models to predict. As the autocompletion tool is not used yet, we have to make some assumptions to get overall performance indicator estimates. We assume that 80% of internet questionnaires will use an enriched job description within the list, through the autocompletion tool, as it occurs in the labor force survey. The remaining 20% do not find what they wanted in the list and declare manually a textual description. Using the Census 2020 paper questionnaires, we are also able to estimate that 35% of the textual descriptions correspond to a job description that appears in the list. Those questionnaires, paper and internet, will be perfectly classified in the 4-digit PCS 2020 class they belong. For the remaining, we should use machine learning model predictions. So, the overall accuracy of the approach

[5] One should keep in mind that the comparison between human annotator coding precision and model accuracy is not completely direct. As the ground truth is deducted from the human annotations, human coding precision obtained here may be slightly overestimated.

can be estimated as

$$p_i = t_{paper,i} \times (.35 + .65 \times p_{ML,i}) + (1 - t_{paper,i}) \times (.8 + .2 \times p_{ML,i}) \qquad (1)$$

where $t_{paper,\,i}$ stands for the share of paper questionnaires for occupation i. Here, we focus on a 4-digit accuracy indicator for salaried occupation and self-employed, *i.e.* observations correctly classified in 1-, 2-, or 3- digit classes or correctly classified as "unclassified" count as errors. This turns to be a 2-digit accuracy indicator for past occupation. These figures are compared to targets that the census team set, based on their expectations and usual quality surveys. These targets were conditions to consider putting the supervised learning approach into production as a complement to autocompletion.

	Current occupation				Past occupation	
	for wage-earners		for self-employed			
	accuracy	test size	accuracy	test size	accuracy	test size
All	88.0 %	5,000	88.7 %	2,000	82.0 %	2,000
Target	> 82 %		> 87 %		> 91 %	
Paper	76.0 %	1,518	71.0 %	621	72.5 %	1,065
Internet	92.9 %	3,482	93.9 %	1,379	92.3 %	935
Target	> 72 %		> 77 %		> 81 %	

Table 7: Combination accuracy: % test sample questionnaires correctly classified at the 4-digit level for salaried and non salaried current occupations and at the 2-digit level for past occupation, with underlying targets.

Table 7 reports results. Overall, the precision estimates at the 4-digit level of the current occupations for wage-earners and self-employed classification exceed the targets, whereas the precision at 2-digit target is not achieved for the past occupation, which counts much more paper questionnaires. The 4-digit precision target is not achieved for the occupation of self-employed when only paper questionnaires are considered. Other breakdown results by main economic activity groups

(5 groups), 2-digit occupational classes, NUTS3, show quite correct results.

In contrast with the current approach for classifying in PCS2003, these levels of precision are achieved without requiring any manual coding during the campaign. However, as shown by the first experiment, models accuracy is expected to decrease year after year if the models are not re-trained regularly.

Using manual coding to gain some more precision and to re-train. With the current system, about 12% of the questionnaires are manually coded. These are the cases when the SICORE rules system does not find a code. In contrast to SICORE, the classifiers trained using supervised learning methods predict always a class, the most probable one. The future should make some productivity gains by using less human coding, but will keep a sizeable part of it for controlling models performance/quality, increasing the classifying quality during the campaign, and re-training.

The last two objectives can be achieved by a kind of active learning approach. Both for regular re-training and for improving the coding quality during the campaign, we plan to have the observations for which the model predictions show the lowest confidence re-coded by humans. The confidence index used to do so is the difference between the probability of the highest probability class and the one of the second highest probability class, see section 3. Doing this, we can expect to gain some accuracy points during the campaign, and at the end, when re-training the models.

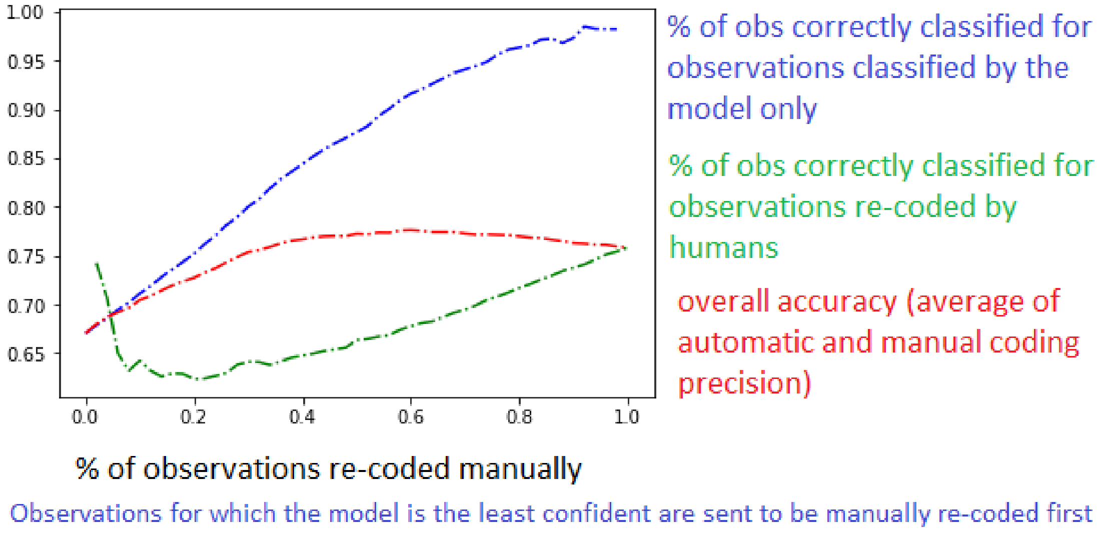


Figure 2: Overall accuracy when observations for which models are the less confident are re-coded manually.

Figure 2 shows the overall accuracy of a process in which observations with the lowest confidence index are gradually send to be re-coded by humans. The overall accuracy increases from 67% (full automatic) to 74% when 40% of the observations (those for which the model is the less confident) are re-coded manually. Even if these observations are also difficult to code by humans, for them, the accuracy of human coding is higher than the model one. Moreover, this combination of automatic and human coding seems to slightly exceed the accuracy of a full human coding. We find similar results and conclusions for the other types of occupations.

Re-training process over a census cycle. A census cycle lasts 5 years. Previous considerations suggest having a regular process of model re-training to take opportunity of human re-coding, campaign after campaign. Model re-training could follow a census cycle:

- Year 1.
  - use the initial train sample (1) to train a first classifier (classifier 1)
  - use classifier 1 to classify year 1 observations, and compute confidence indices

  - have human annotators re-code observations for which model prediction confidence is below a given threshold (which may vary according to available resources) and constitute sample (2)
  - re-train a classifier using sample (1) and sample (2) (classifier 2)
  - several options to chose the final prediction: (1) human coded class when available (sample (2)) and classifier 1 predictions or (2) human coded class when available (sample (2)) and classifier 2 predictions

- Year 2.
  - the train sample becomes sample (1) + sample (2). Use classifier 2 to classify year 2 observations, and compute confidence index
  - have human annotators re-code observations for which model prediction confidence is below a given threshold and constitute sample (3)
  - etc.. proceed as for year 1.
- Year 3 to 5. proceed as for year 2
- Using the last trained classifier to re-classify observations of years 1 to 5 would ensure an homogeneous coding over the census cycle

5. Strategy for production integration

Based on the experimental results presented below, the census team has decided to put into production the new process for the 2024 annual census campaign. During the course of 2022, several working groups bringing together the IT teams, the census, the methodology, and the occupational classification experts plan the production integration strategy, whether on its practical, IT, methodological or organizational aspects. This covers evaluating costs and gains of the integration of (part of) the modules developed during the experiment. This covers defining the new organization of the census production relative to occupational coding, defining different roles and strategy to evaluate and control the quality of coding by the algorithms.

A modular approach The complete classifying pipeline targeted is described in Figure 3. It follows a modular approach and is composed of 6 blocks that can be developed and modified independently.

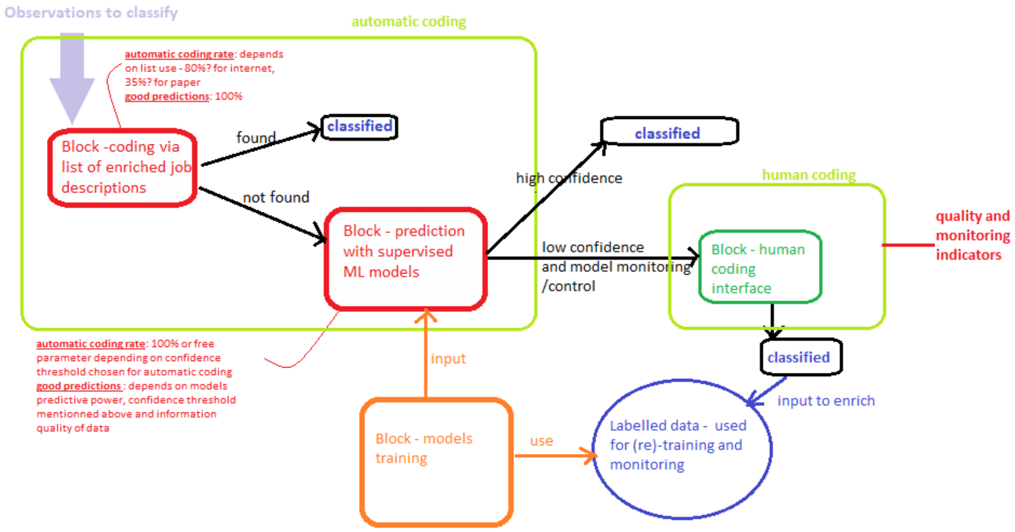

Observations to classify
automatic coding
automatic coding rate: depends on list use - 80%? for internet, 35%? for paper
good predictions: 100%
Block -coding via list of enriched job descriptions
found
classified
not found
Block - prediction with supervised ML models
high confidence
classified
low confidence and model monitoring /control
human coding
Block - human coding interface
quality and monitoring indicators
classified
input to enrich
automatic coding rate: 100% or free parameter depending on confidence threshold chosen for automatic coding
good predictions : depends on models predictive power, confidence threshold mentionned above and information quality of data
input
Block - models training
use
Labelled data - used for (re)-training and monitoring

Figure 3: Census PCS2020 occupational classification strategy

- Block - autocompletion tool and the list of enriched job descriptions, which are mutualized with tools used by the labor force survey.

- Block - ML prediction, which also integrates the preprocessing needed and the service to query the prediction.

- Block - interface for human labeling, based on the current interface and some improvements developed in the interface built for the ad hoc campaign

- Block - model specifying and training, which should be flexible enough to be able to take advantage of ML/IA innovations that could bring more performing models.

- Block- labeled data, which will increase in quality and volume year after year, and are used for training, test, and monitoring

- Block - monitoring indicators

The main issue for the transition to production is to have a prediction tool on a production environment and a strategy selecting observations that will be re-coded by human annotators. The selected scenario consists in providing a web service, using the experimental processing engine developed in Python for the experiment presented here, by integrating it into an architecture allowing the interfacing with client applications, via for example a Web API, also implemented in Python. The IT developments needed (encapsulation, interfacing) are planned to be done in 2023.

The complete specification of the observations selected to be re-coded by human annotators will be done by the census team with the objectives of increasing models accuracy, campaign quality, while remaining within the available resource volumes. To increase the quality of the set of labeled data, it could be chosen to perform double coding + trade-off (as it was done for the *ad hoc* campaign) at the price of having less questionnaires coded. In such a scenario, the occupational experts who took part of the *ad hoc* campaign could be those performing the part of human coding on which the most quality is expected (such as the trade-off part).

Model retraining does not need to be done into a production environment, which offers more flexibility. The complete strategy of re-training will be put in place progressively in particular following the teaching of the first two years, during which the volume of labeled data will rapidly increases.

Organizational aspects. The production integration strategy and first two years of production involves IT teams, census, methodologists, occupational classification experts, each with specific roles. In particular, a datascientist will reinforce the census team in September 2022, with a steering role.

A risk analysis showed that besides IT risks such as the production infrastructure functionalities, and some developing aspects, the main risks concern human resources, datascience experience, and coordination between different actors. Various options, or plans B are planned to deal with these risks.

6. Concluding comments

The implementation strategy of the occupational classification in the census will be refined in 2022 and 2023. Based on this, it will be studied the possibility of having a completely mutualized tool to classify into PCS 2020 data from different sources and for different actors. Indeed, by its volume of data, the annual census survey is particularly adapted to provide voluminous labeled data sets on which models can be trained and test. However, questions of comparability between information from various sources raise. First results suggest that the PCS 2020 model trained on census data does not perform so well in classifying occupations of the labor force survey for instance (around 50%). This a more challenging target.

References

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

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2016): "Enriching Word Vectors with Subword Information," *CoRR*, abs/1607.04606.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019): "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)*, pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joulin, A., E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov (2016): "FastText.zip: Compressing text classification models," *CoRR*, abs/1612.03651.

Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov (2016): "Bag of Tricks for Efficient Text Classification," *CoRR*, abs/1607.01759.

Leroy, T. (2022): “Quelques limites de l’algorithme implémenté dans l’outil Sicore,” in *Proceedings of the Journées de méthodologie statistique de l’Insee.*

Leroy, T., and T. Loisel (2021): “Machine Learning approaches for coding occupations into the new national occupational classification,” in *New techniques and technologies in official Statistics 2021.*

Leroy, T., L. Malherbe, and T. Seimandi (2022): “Application de techniques de machine learning pour coder les profession en PCS 2020,” in *Proceedings of the Journées de méthodologie statistique de l’Insee.*

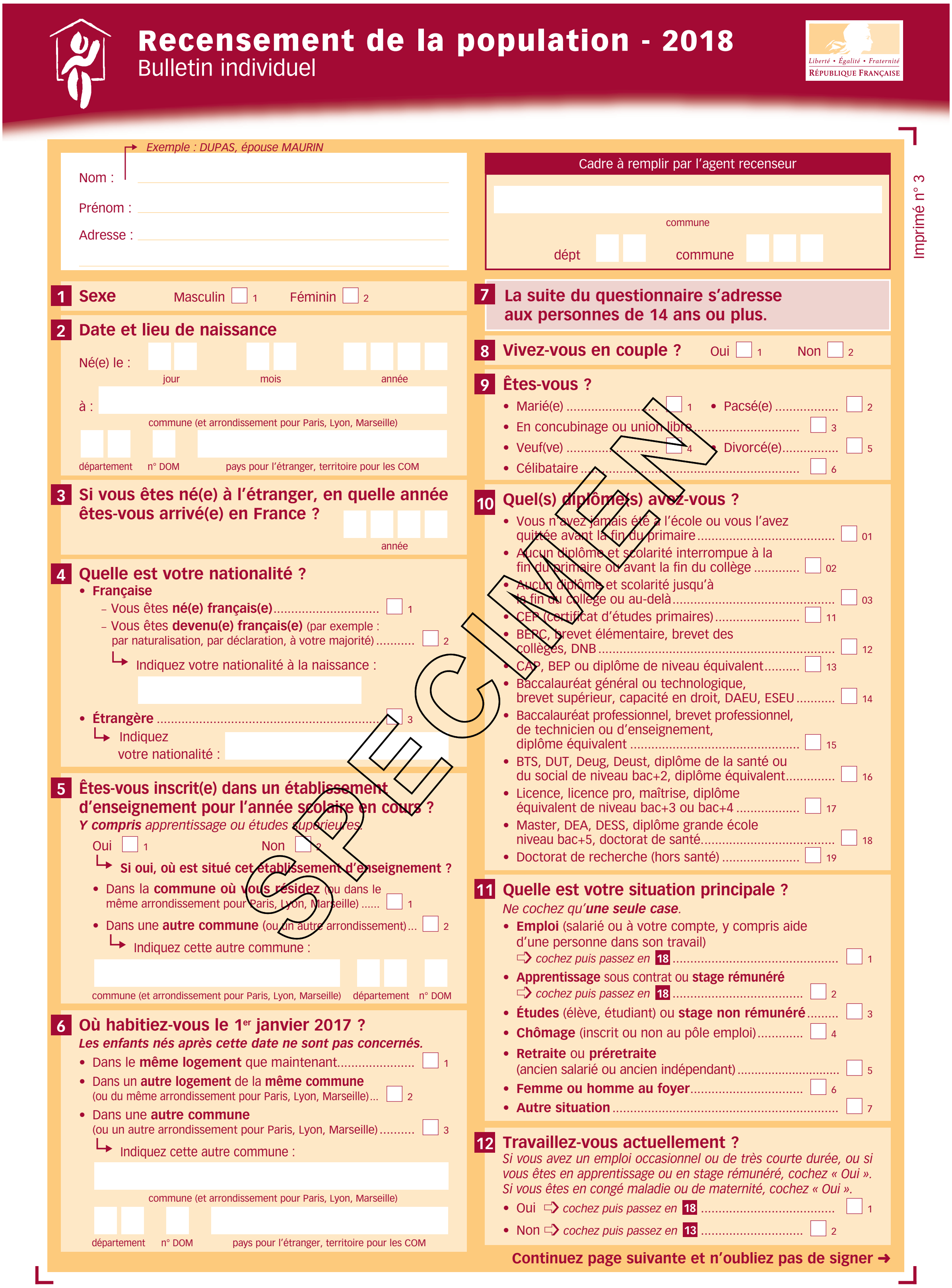
Lorigny, J. (1988): “QUID, une méthode générale de chiffrement automatique,” *Techniques d’enquêtes*, 14.

Martin, L., B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot (2020): “CamemBERT: a Tasty French Language Model,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*

Rivière, P. (1995): “SICORE: système général de chiffrement automatique,” in *Proceedings of the Journées de la méthodologie statistique, INSEE Méthodes n°59-60-61.* INSEE.

Schuhl, P. (1996): “SiCORE, The INSEE Automatic Coding System,” in *Bureau of the Census 1996 Annual Research Conference and Technology Interchange.* Citeseer.

APPENDIX



Recensement de la population - 2018

Bulletin individuel

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Imprimé n° 3

SPECIMEN

Exemple : DUPAS, épouse MAURIN

Nom : ____________________

Prénom : ____________________

Adresse : ____________________

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

Cadre à remplir par l'agent recenseur

____

commune

dépt □□ commune □□□

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

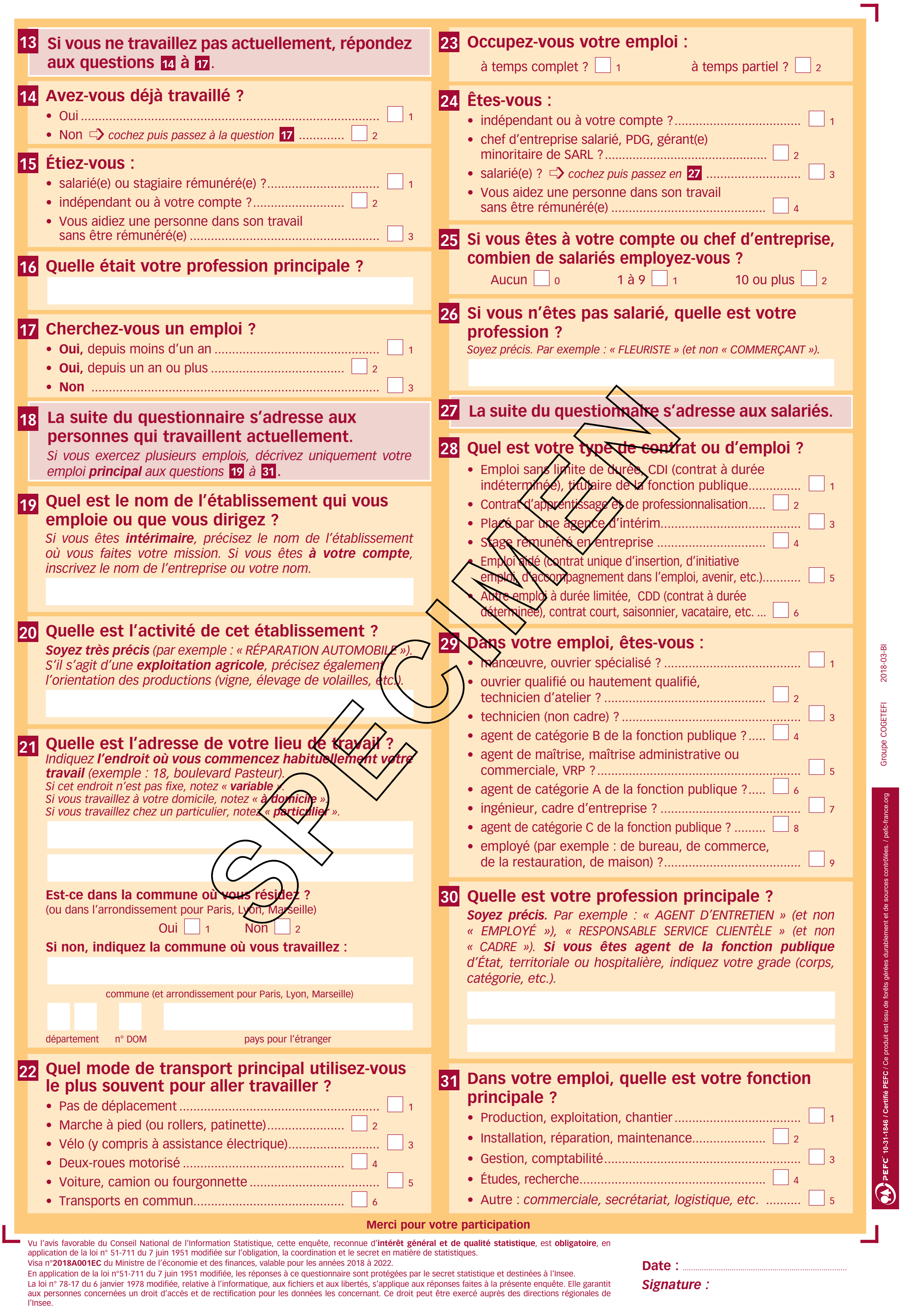
12 Travaillez-vous actuellement ?

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 4: Census questionnaire in 2018 Page 1.



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 et 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

SPECIMEN

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 COGETEFI 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 5: Census questionnaire in 2018 Page 2.