JMS Journées de méthodologie statistique de l'Insee 2018

Analyse des offres d'emploi en ligne : comment coder le métier ?

Version provisoire

Paul ANDREY(*), Maxime BERGEAT(**)

(*) Ensae

(**) Dares, Département métiers et qualifications

paul.andrey@ensae.fr

Mots-clés. Analyse textuelle, Appariement, Données massives, Emplois à pourvoir, *Machine learning*, *Scraping*, Sources multiples

Résumé

Cette communication vise à présenter les travaux expérimentaux effectués par la Dares sur l'analyse des offres postées sur des sites d'offres d'emploi. Dans le cadre du projet européen *ESSNet Big Data – Webscraping job vacancies*, des offres d'emploi publiées en ligne sont collectées de façon automatique avec l'objectif d'améliorer la précision des statistiques sur les emplois à pourvoir grâce à la mobilisation de ces nouvelles données. Les expérimentations réalisées ici consistent à coder le métier (selon la nomenclature des Familles Professionnelles) en fonction du libellé ou du descriptif de l'offre d'emploi. Deux approches sont considérées : l'une mobilisant le descriptif complet de l'offre et des modèles de classification entraînés par apprentissage automatique (*machine learning*), et l'autre se fondant sur une analyse de proximité lexicale (*matching*) entre le libellé de l'offre en ligne et un référentiel d'appellations sur les métiers. Le score F1 des modèles de classification entraînés atteint plus de 80 % pour le niveau le plus agrégé de nomenclature en 22 modalités.

Abstract

This paper aims at presenting experimental work towards auotmated occupation coding for online job ads. Online job ads are collected in order to improve precision and quality of current job vacancy statistics published by the statistical service of the French ministry of work ("Dares"). Job title and job description are the two text variables used here in order to code occupation. We consider two approaches. The first one uses the full job description and is based on machine learning algorithms. The other one is based on a matching between the title of the job ad and a list of job titles compiled by the French National Employment Agency ("Pôle emploi"). The

F1-score of the models trained with machine learning to predict the occupation associated with job ads reaches more than 80 % for the experiments presented in this paper, using the most aggregated level of the occupation nomenclature "Familles Professionnelles", which consists of 22 categories.

Introduction et contexte

La Dares publiait jusqu'en 2017 des statistiques trimestrielles concernant les tensions sur le marché du travail afin d'étudier les inadéquations entre offre et demande de travail selon le métier recherché[1]. En particulier, les métiers étaient identifiés comme en tension à partir de l'étude du *ratio* entre le nombre d'offres d'emploi collectées par Pôle emploi et le nombre de demandeurs d'emploi inscrits à Pôle emploi, ce qui permet de repérer une offre de travail excédentaire et/ou une demande de travail insuffisante. Toutefois, la mesure de la demande de travail des entreprises estimée grâce aux offres collectées par Pôle emploi est incomplète. En effet, tous les projets de recrutement anticipés ne donnent pas lieu à la diffusion d'une offre d'emploi[2], et toutes les offres d'emploi ne sont pas postées sur le site de Pôle emploi par les recruteurs. Par ailleurs, on constate ces dernières années une multiplication des sites proposant des offres d'emploi en ligne, ce qui rend d'autant plus discutable la représentativité des offres collectées par Pôle Emploi, et donc la pertinence de l'exploitation statistique de ces seules offres[6].

Pour étudier sur un champ plus complet les offres d'emploi publiées en ligne, la Dares dispose de deux sources de données potentielles, mobilisées dans cette étude :

— D'une part, dans le cadre de la démarche « Transparence du Marché du Travail » (TMT) initiée par Pôle emploi en 2012, l'opérateur public souhaite utiliser l'ensemble des offres d'emploi disponibles en ligne pour les recentraliser sur son site et donner l'information la plus complète possible aux demandeurs d'emploi. Dans ce cadre, des partenariats bilatéraux ont été conclus entre Pôle emploi et environ 140 sites proposant des offres d'emploi, ces dernières étant rediffusées sur le site de Pôle emploi en tant qu'offres partenaires. Au quatrième trimestre 2017, 55 % des offres diffusées sur le site de Pôle emploi étaient des offres partenaires n'ayant pas été collectées directement par Pôle emploi par ailleurs. Dans le cadre de cette communication, sont mobilisées les offres transmises à Pôle emploi par ses partenaires au cours de l'année 2016, soit un peu plus de 4,3 millions d'offres[10].
— D'autre part, il est possible de récupérer les données brutes sur les sites publiant des offres d'emploi, par collecte automatique de données (*scraping*). Dans le cadre de cette communication, on utilise l'ensemble des offres d'emploi en ligne disponibles à une date de l'été 2017, scrapées à l'aide d'un outil développé spécfiquement à cette fin. Cela représente environ 60 000 offres.

Une fois collectées par *scraping* ou *via* les sites partenaires, les données d'offres d'emploi doivent être structurées pour l'analyse statistique. En particulier, plusieurs variables doivent être codées selon les nomenclatures utilisées, comme le secteur d'activité de l'établissement recruteur ou le métier recherché. Cette publication présente des méthodes et résultats portant sur la codification automatisée du métier selon la nomenclature des Familles Professionnelles (Fap), utilisée à la Dares.

Pour coder le métier, on mobilise ici les données textuelles du libellé et du descriptif des offres d'emploi en ligne. L'intérêt principal de ces deux variables est qu'elles sont, pour la plupart des sites d'offres d'emploi en ligne, disponibles pour toutes les offres. A partir de l'exploitation de ces données textuelles, on teste les performances d'algortihmes de classification à apprentissage supervisé (*machine learning*) et d'appariement d'appellations pour identifier le métier faisant

l'objet de chacune des offres, codé selon la nomenclature Fap. Avec les algorithmes à apprentissage supervisé, pour lesquels on dispose de données permettant leur évaluation, la précision des résultats atteint plus de 80 % pour le niveau agrégé de la nomenclature Fap, comportant 22 modalités.

La section 1 présente les enjeux méthodologiques et techniques associés à l'utilisation des offres partenaires de Pôle emploi et à la collecte automatisée d'offres d'emploi en lignes. Les méthodes adoptées pour le prétraitement des données textuelles en amont de leur exploitation pour la codification du métier y sont également exposées.

La section 2 détaille les méthodes déployées pour codifier le métier associé aux offres étudiées. Les résultats obtenus dans les expérimentations réalisées sont ensuite communiqués.

La section 3 propose quelques éléments de discussion ainsi que des pistes de développement pour des travaux ultérieurs liés à l'exploitation de données non structurées sur les offres d'emploi en ligne.

1 Récupération et structuration des données

Comme indiqué dans la section introductive, deux sources de données sont mobilisées :

- Les offres partenaires diffusées par Pôle emploi. On dispose pour cette étude de données associées à un peu plus de quatre millions d'offres diffusées en 2016.
- Les offres diffusées sur un site d'offres d'emploi en ligne. On dispose pour cette étude de données associées à environ 60 000 offres collectées par *scraping* et représentant l'ensemble des offres en ligne à une même date à l'été 2017.

Ces deux sources de données et les enjeux méthodologiques qui sont associés à leur collecte et à leur exploitation sont présentés tour à tour dans cette section. Pour le *scraping*, des éléments techniques et juridiques sont également exposés.

Une fois collectée, l'information doit faire l'objet d'une structuration. Les données textuelles du libellé et du descriptif de chacune des offres étant au centre des méthodes de codification du métier (voir section suivante), on s'intéresse dans la dernière partie de cette section au prétraitement des données textuelles avant l'étape de codification du métier. Deux méthodes distinctes, employées respectivement pour le libellé et pour le descriptif de chacune des offres, y sont présentées.

1.1 Les offres partenaires de Pôle emploi

Les offres partenaires de Pôle emploi sont l'ensemble des offres transmises à Pôle emploi par des sites de diffusion d'offres d'emploi en ligne contractualisés par l'opérateur public dans le cadre de la démarche « Transparence du Marché du Travail ». Les sites concernés sont de natures diverses, comprenant des sites d'offres d'emploi en ligne sur lesquels postent directement les recruteurs (ou un éventuel intermédiaire), mais aussi des sites dits agrégateurs, qui indexent et agrègent les offres déjà diffusées en ligne par un grand nombre de sites, ainsi que quelques plateformes de recrutement en ligne de grandes entreprises. S'y retrouvent des sites d'importance diverses (en termes de fréquentation et de nombre d'offres diffusées), dont certains visent des segments spécifiques du marché du travail, notamment en termes de secteur d'activité, de qualification des métiers ou de zone géographique.

En amont de leur diffusion, Pôle emploi procède au dédoublonnement des offres partenaires et attribue à chacune un code métier selon la nomenclature du Répertoire opérationnel des métiers

et des emplois (Rome). Une table d'équivalence entre les nomenclatures Fap et Rome, faisant intervenir dans certains cas une variable de qualification, est disponible, et les offres partenaires traitées par Pôle emploi peuvent donc théoriquement être intégrées à l'exploitation statistique existante. Toutefois, pour le moment, une analyse conjoncturelle de ces offres partenaires paraît délicate, du fait de la hausse rapide du nombre de sites partenaires, ce qui limite les possibilités de construire un champ constant pour l'étude des offres diffusées par Pôle emploi.

En revanche, de telles données, où sont à la fois disponibles des informations sur le descriptif de l'offre et un code métier attribué par les services de Pôle emploi, peuvent être utiles pour mettre en place et évaluer des méthodes de codification du métier. Elles peuvent en effet servir de données d'évaluation pour les algorithmes d'appariement (voir section 2.1), ainsi que de données d'entraînement pour l'apprentissage supervisé de la tâche de classification des offres (voir section 2.2).

Dans le cadre de cette étude, les titres des offres partenaires ne figurent pas dans les données, bien que Pôle emploi les ait utilisés pour associer un code Rome à chaque offre, selon une méthode similaire à celle présentée à la section 2.1. Nous ne pouvons donc ni évaluer nos algorithmes d'appariement sur ces données, ni évaluer la qualité des traitements réalisés par l'opérateur public, de laquelle dépend pourtant la pertinence de l'utilisation de ces données pour entraîner de manière supervisée des algorithmes de classification. Par ailleurs, les descriptifs des offres sont tronqués au-delà de 1024 caractères, cela est susceptible de légèrement dégrader la qualité de ce jeu de données.

1.2 *Scraping* de sites d'offres d'emploi

Le *scraping* consiste à extraire et à structurer automatiquement une information véhiculée par un système informatique produisant des sorties lisibles par un humain. Autrement dit, il s'agit d'automatiser la collecte d'une information déjà accessible « visuellement ». On parle alors de *web scraping* pour désigner la collecte automatisée d'informations publiquement accessibles sur internet. Dans le cas des offres d'emploi en ligne, une telle pratique permet d'envisager la collecte et la structuration d'informations obtenues à partir d'un ensemble de sites constituant un panel représentatif pour les offres d'emploi diffusées en ligne. Toutefois, le grand nombre d'acteurs en ligne relègue l'exhaustivité au rang d'idéal, et pose la question de la représentativité des offres effectivement collectées. Dans le cadre de l'étude de l'apport des offres d'emploi en ligne pour mesurer les tensions sur le marché du travail, et en collaboration avec d'autres pays européens au sein du groupe de travail *Webscraping job vacancies* du projet *ESSnet Big Data* de la Commission Européenne, un outil de *scraping* d'offres d'emplois en ligne a été développé, en langage Python.

Le *scraping* d'une page web consiste à effectuer une requête HTTP afin d'obtenir le code source (en HTML) de la page, puis à extraire l'information pertinente de ce code source au moyen d'un parseur (analyseur sémantique). Afin de collecter les données d'offres d'emploi sur un site à une date donnée, l'outil parcourt les URLs (déterministes) des pages indexant les offres en ligne sur le site (typiquement, les pages de résultats d'un moteur de recherche interne au site listant les offres d'emploi, sans critère de recherche), afin d'en extraire la liste exhaustive des URLs (adresses web) des pages détaillant chacune des offres. L'outil peut alors parcourir les URLs collectées afin de collecter et structurer les informations associées à chaque offre. L'outil développé comprend une structure générale de fonctionnement, à laquelle s'articulent pour chacun des sites visés deux parseurs spécifiques (l'un pour les pages indexant les offres, l'autre pour les pages des offres elles-mêmes). Il est également possible de définir la fréquence des requêtes afin de ne pas surcharger les serveurs des sites qu'on veut scraper.

La législation sur le *scraping* et l'exploitation de données collectées par ce moyen s'avère ambivalente, n'interdisant ni n'autorisant formellement ces pratiques. Un groupe de travail européen de l'*ESSnet Big Data* a publié au premier semestre 2017 un rapport concluant que l'utilisation par la statistique publique de données scrapées ne semblait pas poser problème, tout en préconisant des règles de bonne conduite incluant le fait de prévenir les propriétaires des sites ciblés, de respecter leur opposition le cas échéant et d'adopter les limites spécifiés par les sites en termes de fréquences de requêtes[5].

Du fait de la fréquence imposée des requêtes, le *scraping* tend à prendre beaucoup de temps, la plupart des sites demandant un temps d'attente de deux à cinq secondes entre chaque requête, ce qui excède largement le temps nécessaire pour conduire une requête et parser le code source reçu. De plus, un site peut toujours refuser les requêtes entrantes, notamment lorsque des requêtes répétées sont conduites, ce qui remet en cause la stabilité du *scraping* pour collecter des données. Enfin, les sites voyant l'architecture de leurs pages évoluer, les parseurs spécifiques à chacun d'entre eux requièrent un travail de maintenance régulière pour poursuivre une collecte régulière de données dans le temps. Pour toutes ces raisons, le *scraping* apparaît comme une méthode imparfaite d'accès à des données, et rend la contractualisation de sites (à la manière de Pôle emploi) plus intéressante. Le *scraping* peut néanmoins servir de complément aux offres partenaires diffusées sur le site de Pôle emploi, permettant de cibler les sites considérés comme pertinents et refusant un partenariat. La question qui se pose alors est celle du choix des sites ciblés : on peut notamment se demander s'il est préférable de viser des sites généralistes, des sites spécialisés, des agrégateurs ou encore des sites d'entreprises en vue de constituer un échantillon représentatif et stable dans le temps pour les offres d'emploi diffusées en ligne.

Dans le cadre de cette publication, 60 000 offres scrapées sur un même site à une même date sont exploitées. Le titre et le descriptif de chacune de ces offres a été collecté, ainsi que l'ensemble des informations structurées disponibles, qui varient en nombre et en précision selon les offres (qualification et expériences requises, localisation géographique, identité du recruteur, secteur d'activité, date de première publication et de dernière mise à jour de l'offre...).

1.3 Prétraitement des données textuelles

Avant d'utiliser les informations textuelles contenues dans le libellé et le descriptif des offres d'emploi, un nettoyage des données est opéré. Outre l'harmonisation des caractères (encodage, ponctuation et caractères spéciaux, dont accents) et le retrait d'un ensemble déterminé de mots fréquents ne convoyant pas d'information pertinente (connecteurs logiques, pronoms, *etc.*), il s'agit de lemmatiser chacun des mots. La lemmatisation consiste à rattacher toutes les formes fléchies d'un même terme à un même lemme, retirant ainsi le contexte grammatical du terme ; par exemple, les mots « vendeurs » et « vendeuse » seront tous deux remplacés par le lemme « vendeur ». Le choix du lemme comme unité d'analyse vient du fait que les algorithmes utilisés pour la codification du métier à partir du texte se basent soit sur la comparaison de termes deux-à-deux, soit sur des vecteurs synthétisant un texte sans retenir l'ordre des termes. Dans le premier cas, la lemmatisation permet un meilleur rapprochement de certains mots (déjà ramenés à une forme identique) ; dans le second, elle réduit la dimension du vocabulaire et donc des vecteurs employés, *a priori* sans perte d'information pertinente.

Dans le cadre de cette étude, la lemmatisation n'est pas réalisée de la même façon pour le libellé (titre) et le descriptif des offres d'emploi, en raison de la nature différente de ces textes et de leur traitement différencié pour la codification du métier.

Le titre des offres est généralement un groupe nominal désignant le poste faisant l'objet d'une annonce et pouvant inclure des indications concernant le recruteur ou le profil du candidat,

parfois précédé du terme « recherche » (« vendeur en CDI » / « recherche vendeur en CDI ») et plus rarement d'autres éléments. Ici, la lemmatisation est réalisée de manière indépendante pour chaque mot, à l'aide du corpus lexical Morphalou. Morphalou est un lexique de formes fléchies du français conçu par des chercheurs de l'ATILF (Université de Nancy - CNRS) distribué librement au format XML. Le corpus propose initialement pour un grand nombre de lemmes un ensemble de formes fléchies ; en inversant ce dictionnaire, on peut donc rattacher tout terme contenu dans le lexique à un lemme. Pour les formes fléchies homonymes, un lemme nominal est préféré à un lemme verbal, car *a priori* plus probable dans un titre d'offre. Les termes n'apparaissant pas dans le lexique sont eux conservés tels quels [1].

Le descriptif de l'offre d'emploi consiste en un paragraphe plus structuré, comprenant plusieurs phrases. Ici, la lemmatisation est effectuée en tenant compte du contexte des mots afin d'inférer leur fonction grammaticale et donc de résoudre le problème des homonymies. Le terme « recherche » sera ainsi alternativement rattaché à son lemme nominal (« recherche ») ou verbal (« rechercher »), selon la fonction grammaticale identifiée. Cela pourra permettre par exemple d'attribuer un lemme différent au terme « recherche » dans les phrases « notre entreprise recherche un boulanger » et « notre entreprise recrute un ingénieur en recherche et développement ». On utilise ici le programme TreeTagger. TreeTagger est un outil d'annotation automatique de texte visant à identifier la fonction grammaticale des termes (*part-of-speech tagging*), développé par un chercheur de l'université de Stuttgart, et utilisable gratuitement pour des fins de recherche. L'outil couple un lexique pré-appris de termes et formes fléchies avec un algorithme d'inférence analysant le texte par trigrammes (triplets de termes). Dans cette étude, on utilise le modèle pré-entraîné pour le traitement de la langue française distribué avec TreeTagger, et le lemme identifié par l'outil pour chacun des termes lui est substitué. Les termes inconnus de TreeTagger sont, eux, conservés tels quels [2].

2 Codification du métier

Dans cette publication, on s'intéresse exclusivement à la codification du métier pour les offres d'emploi en ligne collectées. Deux types de méthodes s'appuyant respectivement sur le libellé et le descriptif de l'offre, qui ont été préalablement nettoyés et lemmatisés, sont testés. La première approche consiste à apparier le libellé de chaque offre à un référentiel d'appellations de métiers, en l'occurrence le Rome de Pôle emploi. La seconde approche repose sur l'apprentissage automatique (*machine learning*) de la tâche de classification d'une offre dans un métier à partir de son descriptif selon un modèle statistique. Ces deux approches font l'objet des deux premières sous-sections de cette partie ; y sont présentés leur principe général et les modalités pratiques testées pour cette étude. Une dernière sous-partie fait état des résultats obtenus.

La nomenclature utilisée pour codifier le métier est celle des Familles Professionnelles (Fap). Cette nomenclature, utilisée pour les études sur les métiers menées à la Dares, peut être obtenue à la fois à partir du Rome de Pôle emploi et de la nomenclature des Professions et Catégories Socioprofessionnelles (PCS). Il s'agit d'une nomenclature hiérarchique à trois niveaux, comportant respectivement 22, 87 et 225 modalités.

1. On pourra trouver plus d'informations sur le lexique Morphalou ici.
2. On pourra trouver plus d'informations sur l'outil Treetagger sur ce lien.

2.1 Méthode d'appariement

L'approche par appariement repose sur la comparaison entre le libellé de chaque offre et un référentiel comportant des appellations de métiers et, pour chacune d'elles, un code métier associé. Il s'agit pour chaque offre d'identifier l'appellation de référence concordant le mieux (ou d'identifier qu'aucune appellation de concorde), afin de lui attribuer le code métier associé à cette appellation. Cette méthode est celle utilisée par Pôle emploi afin d'attribuer un code Rome à chacune des offres partenaires, bien que les détails de l'implémentation diffèrent de ceux utilisés pour cette étude.

Cette méthode fait l'hypothèse que le métier est explicitement désigné dans le libellé de chacune des offres, et repose sur le choix d'un référentiel d'appellations de métiers ainsi que d'une fonction de distance pour quantifier la similarité entre le libellé nettoyé et chacune des appellations du référentiel. Au terme de la procédure d'appariement, une offre se voit non seulement associée un code métier, mais aussi un score de similarité avec l'appellation de référence attachée à ce code, qui peut dans l'idéal servir d'indicateur de confiance dans la validité de cet appariement, quoi que l'interprétation et la pertinence de cet indicateur peut varier selon la fonction de similarité considérée. Un algorithme peut ainsi être entièrement défini *a priori* par le choix d'une certaine fonction de similarité, mais il est également possible de comparer les performances de différentes fonctions ou de différents paramétrages de celles-ci sur un jeu de données supposé représentatif et étiqueté de manière fiable.

Dans la pratique, on a ici utilisé comme référentiel les appellations métier du Répertoire opérationnel des métiers (Rome) de Pôle emploi. Il s'agit d'environ 11 000 appellations de métiers, dont chacune est associée à l'un des 531 codes Rome en considérant la granularité la plus fine de la nomenclature Rome. Il est ensuite possible de convertir le code Rome en Fap, bien qu'en toute rigueur cette conversion devrait faire intervenir une variable de qualification du poste pour atteindre sa précision maximale. Chacune des appellations du référentiel a fait l'objet de la même procédure de nettoyage et de lemmatisation que les libellés des offres que l'on cherche à apparier, afin d'éviter de créer artificiellement des disparités entre textes liées à la lemmatisation.

La fonction de distance utilisée pour cette étude est de la forme suivante, qui est celle à laquelle Pôle emploi a recours :

Soit deux textes A et B, considérés comme des ensembles de mots.

$$distance(A,B) = \frac{\alpha \times demi\text{-}distance(A,B) + (1-\alpha) \times demi\text{-}distance(B,A)}{2}$$

$$demi\text{-}distance(A,B) = \frac{\sum_{a \in A}(\exists b \in B, a \sim b)}{Card(A)}$$

avec $\sim$ une notion d'équivalence entre deux mots.

On compte donc la proportion de mots de l'offre d'emploi dont il existe un mot « équivalent » dans le libellé du référentiel d'appellations, et vice-versa. On utilise les deux demies-distances afin de ne pas accorder un score trop important si les libellés à comparer contiennent un nombre différent de lemmes. Les tests réalisés dans le cadre de cette étude utilisent la valeur $\alpha = 0.5$, pondérant équitablement les deux demi-distances, ce qui a probablement à voir avec le prétraitement des données textuelles.

Pour cette étude, on définit l'équivalence entre deux mots comme le fait d'obtenir un score de similarité au sens de la distance de Jaro-Winkler supérieur à un certain seuil, dont la valeur choisie parmi celles testées est 0.9. Le choix de cette distance parmi de multiples distances d'édition vient de sa valorisation des termes possédant les mêmes quatre premiers caractères et de sa faible pénalisation de mots faisant l'objet de fautes d'orthographe simples (insertion ou inversion

de caractère). Elle semble donc particulièrement adaptée pour rapprocher deux lemmes de même racine, ou un terme pour lequel la lemmatisation a échoué du fait d'une erreur orthographique, du lemme auquel il relève. La fonction « d'équivalence » ainsi définie peut permettre de considérér des termes comme équivalents malgré de légères différences orthographiques, tout en demeurant un critère assez strict de similarité.

Pour optimiser le coût computationnel de l'appariement, la précédente fonction de distance entre appellations n'est appliquée que pour les couples de libellé d'offre et d'appellation de référence tels qu'il existe au moins un lemme apparaissant à l'identique dans ces deux textes. Ceci permet d'éviter de coûteux calculs de distance entre mots pour des appellations pour lesquelles, selon toute vraisemblance, l'algorithme ne pourrait aboutir, ou aboutirait à une solution peu fiable (sauf cas extrêmes envisageables).

2.2 Méthode de classification à apprentissage supervisé

L'approche par apprentissage supervisé (*machine learning*) repose sur la définition d'une tâche de classification par un modèle statistique. Cette tâche consiste, à partir d'un vecteur numérique représentant le descriptif d'une offre, à prédire la Fap du métier de l'offre. Les paramètres du modèle sont alors appris automatiquement à partir d'un jeu de données d'entraînement, comportant la variable cible à prédire. Une fois la phase d'apprentissage passée, il est possible d'évaluer les performances du modèle sur des données de test, afin d'évaluer sa capacité à classifier correctement les offres. Par la suite, le modèle peut être utilisé pour classifier des données pour lesquelles la variable cible n'est pas disponible, avec un taux de succès supposé similaire à celui précédemment évalué, sous réserve que ces données soient sujettes à des relations entre leurs valeurs et la variable cible similaires à celles présentes dans les données d'entraînement.

La représentation du descriptif des offres en vecteurs numériques requiert que soit arrêté un vocabulaire de taille fixe, ici notée N. Un descriptif est alors représenté par un vecteur dans $\mathbb{R}^N$ tel que sa i-ème valeur indique la présence du i-ème mot du vocabulaire dans le descriptif. Ces valeurs peuvent par exemple consister en une indicatrice de présence du terme dans le descriptif (le vecteur est alors un vecteur de $\{0,1\}^N$, ou en un décompte du nombre d'occurrences du terme dans le descriptif. Les résultats présentés par la suite ont eux été obtenus en réduisant chaque descriptif aux termes appartenant à un vocabulaire donné (voir paragraphe suivant) et en calculant pour chacun de ces termes un score TF-IDF (*term frequency - inverse document frequency*). Ce score est égal, pour un terme donné, au quotient entre la part des termes du descriptif (réduit) égaux à ce terme et le logarithme de la part des descriptifs parmi le corpus d'offres traité contenant ce terme. Dans la pratique, le dénominateur de ce quotient est calculé pour un premier corpus de documents (à partir duquel est également sélectionné le vocabulaire considéré), et réemployé à l'identique lors du calcul de scores sur tout autre document par la suite. Le score TF-IDF est utile pour surpondérer les mots d'un descriptif qui sont rarement présents pour les autres offres d'emploi (mots potentiellement très spécifiques d'une catégorie de métiers).

Comme indiqué à la section 1.3, les descriptifs des offres font l'objet d'un nettoyage textuel et d'une lemmatisation, ce qui concourt à réduire la taille du vocabulaire total au sein du premier corpus d'offres considéré. Afin de sélectionner un vocabulaire restreint aux termes convoyant le plus d'information *a priori* pertinente, et dont la dimension ne rende pas le coût computationnel de l'entraînement de modèles démesuré, deux approches ont été comparées, dont la seconde s'est avérée la meilleure et a produit les résultats présentés par la suite. La première consiste à conserver les termes apparaissant *a minima* dans un certain nombre de descriptifs parmi ceux du corpus, ici fixé à 0.5 % de la taille du corpus. La seconde consiste à conserver les termes qui, pour l'une au moins des modalités de la variable cible, apparaissent *a minima* dans une certaine part

des descriptifs appartenant à cette modalité, ici fixée à 2%. Cette seconde méthode appliquée aux offres partenaires de Pôle emploi a permis de sélectionner un vocabulaire d'environ 1 500 termes.

Les données ici utilisées pour entraîner différents modèles par apprentissage automatique supervisé et pour évaluer ceux-ci sont les offres partenaires de Pôle emploi sur l'année 2016, qui comportent un code Rome attribué par Pôle emploi, que l'on a converti en Fap. Il est à noter qu'est alors faite l'hypothèse de la validité des codes Rome présents dans les données et obtenus par les équipes de Pôle emploi, à partir d'une méthode d'appariement non évaluée dans ce papier.

Dans le cadre de cette étude, plusieurs modèles pouvant apprendre une tâche de classification multinomiale ont été entraînés de manière supervisée, dont trois voient leurs résultats présentés par la suite :

— **Régression logistique**. Une régression logistique est un modèle linéaire, dont les coefficients sont estimés par maximisation de la vraisemblance. Dans sa forme élémentaire, ce modèle apprend une tâche de classification binaire et retourne une probabilité d'appartenance à l'une des classes. Il est possible d'utiliser un modèle polytomique afin d'apprendre une tâche de classification multinomiale, cependant il est apparu pour cette étude que de meilleurs résultats étaient obtenus en combinant des modèles binomiaux prédisant pour chaque modalité la probabilité d'appartenir à celle-ci plutôt qu'à toute autre.
— **Forêts aléatoires**. Une forêt aléatoire est composée d'un ensemble d'arbres de décision, entraînés de manière ensembliste. Chacun des arbres est appris automatiquement à partir d'un échantillon aléatoire des données d'entraînement, tiré avec remise pour les offres utilisées et sans remise pour les composantes des vecteurs d'entrée retenues. L'utilisation de ce *bootstrap* permet d'obtenir des arbres non colinéaires, dont l'agrégation des prédictions permet de rendre les résultats plus robustes.
— **Perceptron multicouche**. Un perceptron multicouche est un type de réseau de neurones artificiels constitué d'une couche d'entrée, d'une ou plusieurs couches cachées et d'une couche de sortie, toutes pleinement connectées (*i.e.* chaque unité d'une couche reçoit en entrée les sorties de chacune des unités de la couche précédente, et alimente chacune des unités de la couche suivante). Chaque unité comporte un ensemble de poids permettant de transformer linéairement ses entrées, par la suite passées à une fonction dite fonction d'activation retournant la sortie de l'unité. La couche de sortie comporte autant d'unités que de modalités de la variable cible, dont les sorties sont normalisées par la fonction softmax pour sommer à un et ainsi constituer une distribution de probabilités d'appartenance à chacune des modalités. Les poids sont appris par rétro-propagation d'une fonction de perte, qui est ici l'entropie croisée des probabilités prédites.

Pour chacun de ces modèles, un certain nombre de paramètres permettant de spécifier le modèle et/ou sa procédure d'apprentissage, dits hyperparamètres, ont fait l'objet d'un travail d'optimisation (*tuning*) par méthode dite de *grid search*. Il s'agit, pour l'ensemble des hyperparamètres considérés pour un modèle, de fixer un nombre fini de valeurs envisagées, puis, pour chacune des combinaisons de valeurs possibles, d'évaluer les performances atteignables toute choses égales par ailleurs par un modèle spécifié avec ces valeurs. On a adopté pour faire cela une procédure de validation croisée, consistant à partitionner les données d'entraînement en k groupes (ici $k = 5$) de taille égale et à entraîner, pour chaque spécification de modèle envisagée, k modèles en prenant à chaque fois l'un des groupes comme données de validation et les $k - 1$ autres comme données d'entraînement. Cette procédure vise à limiter les risques de surapprentissage, en limitant le risque que les performances atteintes par un modèle soient liées à la découpe spécifique des données entre jeux d'entraînement et de validation.

— **Régression logistique**. Il est possible d'ajouter un terme de pénalisation à la fonction de vraisemblance du modèle, afin de limiter le nombre de variables à utiliser lorsqu'elles sont redondantes. On a ici testé l'absence de pénalisation et les pénalisations par norme L1 (dite méthode du lasso) et L2, en cherchant pour chacune de ces dernières normes le paramètre de poids de la pénalisation optimal.
— **Forêts aléatoires**. On a ici fait varier le nombre d'arbres composant la forêt (50, 100, 200 ou 500), le nombre minimal d'observations par nœud terminal (feuille) de chaque arbre (1, 10 ou 100) ainsi que le nombre de variables que l'arbre peut utiliser dans les critères de décision établis parmi celles disponibles à l'entraînement (100 variables, un nombre égal à la racine carrée du nombre de variables disponibles, ou autant de variables que jugé optimal par l'algorithme).
— **Réseau de neurones**. Plusieurs fonctions d'activation ont été comparées (fonction identité, i.e. pas de fonction d'activation ; fonction logistique ; fonction de rectification (relu)), ainsi que l'ajout d'une seconde et troisième couches de 100 unités à un modèle initial comprenant une unique couche cachée de 100 unités.

Différentes combinaisons des paramètres ont été testées, et les résultats présentés dans la sous-section ci-dessous correspondent aux meilleurs résultats obtenus en termes de pouvoirs prédictifs. Ces résultats pourraient être consolidés en allant plus loin dans le test autour des différents paramètres (*tuning*).

2.3 Résultats

Pour la méthode fondée sur l'appariement du libellé du titre de l'offre d'emploi, il n'est pas possible à l'heure actuelle de fournir une évaluation quantitative de la qualité de l'appariement du fait du manque de données disponibles pour constituer un jeu de données d'évaluation. En effet, on ne dispose pas pour le moment de données contenant à la fois le libellé d'offres d'emploi et un code Rome considéré comme fiable (voir section 1.1). Par conséquent, les résultats et visuels proposés dans la suite de cette section ne concernent que les prédictions réalisées avec des modèles à apprentissage supervisé. On dispose en effet pour celles-ci de données pouvant servir à l'évaluation des résultats, les offres partenaires (voir, à nouveau, la section 1.1).

On présente donc désormais les résultats pour les méthodes prédictives à partir des descriptifs des offres d'emploi sur les données des offres partenaires de Pôle emploi. On dispose pour ces données à la fois de la variable à prédire (le métier recherché, codé par les services de Pôle emploi) et du descriptif de chaque offre. Pour évaluer la performance de la codification, on sépare donc l'échantillon utilisé en 2 parties avec :

— Un **échantillon d'apprentissage** utilisé pour entraîner les modèles de *machine learning* utilisés. Il est ici constitué de 500 000 offres choisies aléatoirement parmi les données disponibles.
— Un **échantillon test** dont on se sert pour évaluer la performance de chacun des modèles entraînés. On compare alors la modalité prédite par le modèle pour chaque offre avec la modalité de la variable métier codée par Pôle emploi et retranscrite dans la nomenclature Fap.

On peut détailler les résultats en fonction des modalités des métiers, en distinguant notamment deux types d'indicateurs pour chacune des 22 modalités à prédire, aggrégés par un troisième :

— La **précision**, qui correspond à la part de documents correctement classés parmi l'ensemble des documents codés selon un métier recherché.

— Le **rappel**, qui représente la proportion de documents retrouvés parmi ceux possédant initialement la modalité correspondante.
— Le **score F1**, qui est défini pour chaque catégorie comme la moyenne harmonique de la précision et du rappel :

$$Score\ F1\ = \frac{2}{\frac{1}{rappel} + \frac{1}{précision}} = 2 \times \frac{précision \times rappel}{précision + rappel}$$

Par ailleurs, des indicateurs globaux sont ensuite calculés. Il s'agit des indicateurs précédents moyennés par catégorie à prédire, en fonction du nombre d'observation appartenant réellement à la catégorie concernée dans les données initiales. La précision globale, le rappel et le score F1 globaux du modèle sont donnés dans la suite de ce papier pour l'évaluation et la comparaison des modèles.

La table 1 présente les scores obtenus par trois modèles distincts, introduits à la section 2.2, laquelle détaille également le type de représentation des descriptifs d'offres utilisés par ces modèles. Les hyperparamètres considérés pour les modèles ont été sélectionnés selon la méthodologie de recherche sur une grille de paramètres (*grid search*) et de cross-validation précédemment introduite. Pour la régression logistique, une pénalisation par norme L1 des poids du modèle est utilisée, avec un coefficient de pénalisation fixé à 5. La forêt aléatoire est, elle, composée de 200 arbres dont les branches sont ramifiées jusqu'à ne déboucher que sur un unique cas lors de leur entraînement, et libres d'utiliser autant de variables que jugé optimal parmi celles dont ils disposent. Quant au perceptron multi-couches, on a constaté que l'ajout de couches cachées additionnelles rendait instable sa convergence tandis que les résultats obtenus après un maximum de 200 étapes d'entraînement ne montraient pas de variation marquée, d'où l'emploi d'une unique couche cachée de cent unités ; la fonction d'activation choisie pour ces unités est, elle, la fonction logistique. On note enfin que pour tous les modèles entraînés au cours de l'affinage des hyperparamètres ont produit des résultats similaires sur chacun des blocs de la validation croisée, ne faisant donc pas état de surapprentissage.

Méthode	Précision globale	Rappel global	Score F1 global
Régression logistique	78.6 %	79.9 %	79.3 %
Forêts aléatoires	80.3 %	80.1 %	80.2 %
Réseau de neurones	80.5 %	81.2 %	80.8 %

TABLE 1 – Performances pour la prédiction de trois méthodes pour coder le métier recherché

On constate que les résultats sont globalement encourageants : en moyenne, 4 offres d'emploi sur 5 sont correctement classifiées, et ce malgré le nombre relativement important de catégories à prédire, et la proximité des scores de précision et de rappel montrent l'absence de biais systématique vers les faux positifs ou les faux négatifs. Les résultats pour les taux de précision et de rappel globaux sont relativement proches, quelle que soit la méthode utilisée pour la prédiction. La mobilisation d'un réseau de neurones conduit toutefois à un pouvoir prédictif en moyenne légèrement supérieur, malgré un coût computationnel très supérieur par rapport aux méthodes utilisant des régressions logistiques ou, dans une moindre mesure, des forêts aléatoires [3] Les résultats présentés ici sont cohérents avec les expérimentations réalisées dans d'autres pays européens, notamment en Allemagne et en Belgique dans le cadre du projet européen {ESSNet Big Data - Webscraping Job vacancies}[5]. En Allemagne par exemple, les tests réalisés pour coder le secteur d'activité des entreprises postant des offres d'emploi indiquent que la méthode qui permet d'atteindre le

3. Sur la même machine, plusieurs heures ont été nécessaires pour l'entraînement du réseau de neurones, contre une heure environ pour la forêt aléatoire et quelques minutes pour la régression logistique.

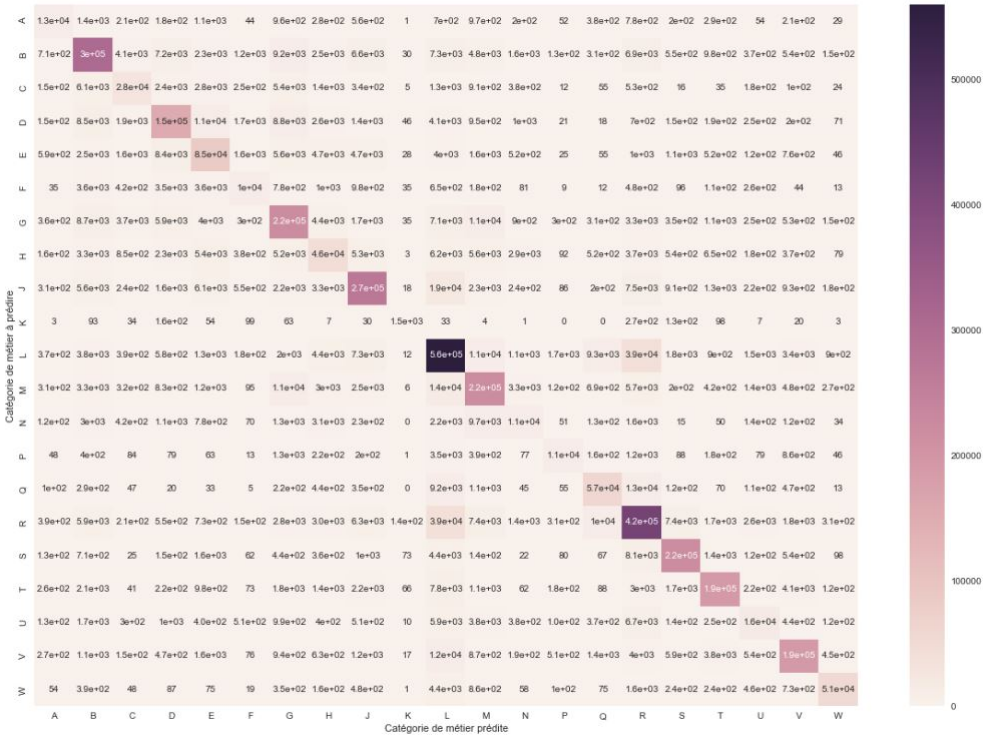

FIGURE 1 – Matrice de confusion obtenue avec un réseau de neurones

meilleur taux de précision globale est la méthode de la classification naïve bayésienne[8], qui est très rapide pour réaliser l'estimation. De même, la précision du codage réalisé est sensiblement similaire quelle que soit la technique utilisée pour la prédiction. Le codage avec la classification naïve bayésienne n'a pas été testé au cours de la présente étude.

Pour préciser les résultats, il est également possible d'étudier les prédictions réalisées catégorie par catégorie. Pour effectuer cela, il est usuel de considérer la matrice de confusion. Cette matrice carrée qui comporte autant de lignes que de modalités à prédire permet de mieux comprendre les erreurs de prédiction réalisées : quelles sont les catégories les plus touchées, et quelles catégories de métier sont alors indûment prédites. Les résultats obtenus en utilisant le perceptron multi-couches sont désormais détaillés.

La figure 1 présente la matrice de confusion obtenue en classifiant le métier recherché dans les offres d'emploi avec un réseau de neurones. Les résultats ne sont ici pas normalisés selon le nombre d'offres par catégorie de métier. On peut pour préciser considérer des matrices normalisées comme présentées dans les figures 2 et 3. Cela permet d'étudier respectivement les erreurs de prédiction réalisées en termes de précision et de rappel en fonction des métiers des offres d'emploi de l'échantillon test. Les résultats laissent apparaître de fortes différences selon les métiers. On constate par exemple que c'est le cas de la catégorie de métier N (personnels d'études et recherche) qui est difficile à prédire correctement (précision de 43 % et rappel de 28 % pour les offres d'emploi concernant les métiers de cette catégorie). Les offres d'emploi pour ces métiers sont souvent (dans 28 % des offres d'emploi pour des métiers d'études et de recherche, d'après la figure 3) considérés à tort comme des offres pour les métiers de la catégorie M (métiers de l'informatique et des télécommunications). De même, les offres d'emploi de l'échantillon test catégorisées comme des métiers d'études et de recherche concernent souvent en réalité des métiers de l'informatique et des télécommunication (catégorie M) ou des offres pour les ingénieurs et cadres de l'industrie (catégorie H), dans respectivement 13 % et 11 % des cas (figure 2). Pour certains métiers comme

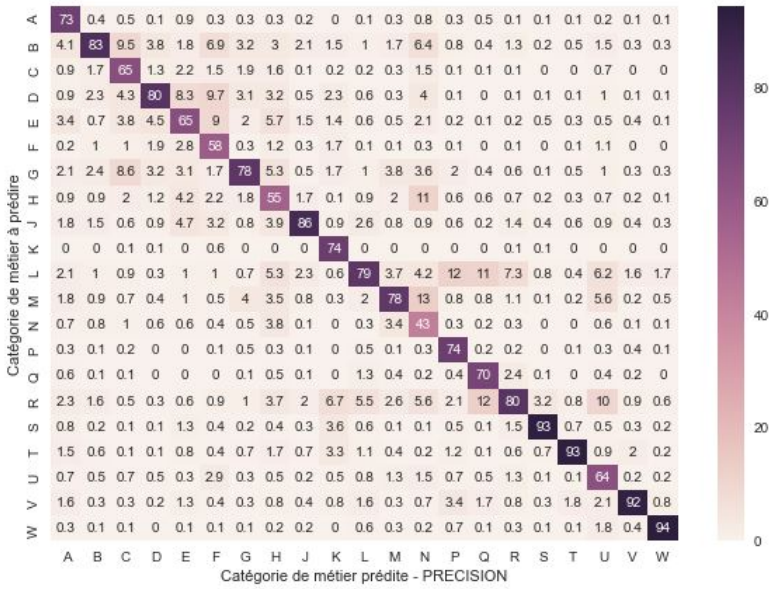

FIGURE 2 – Matrice de confusion normalisée pour le calcul de la précision

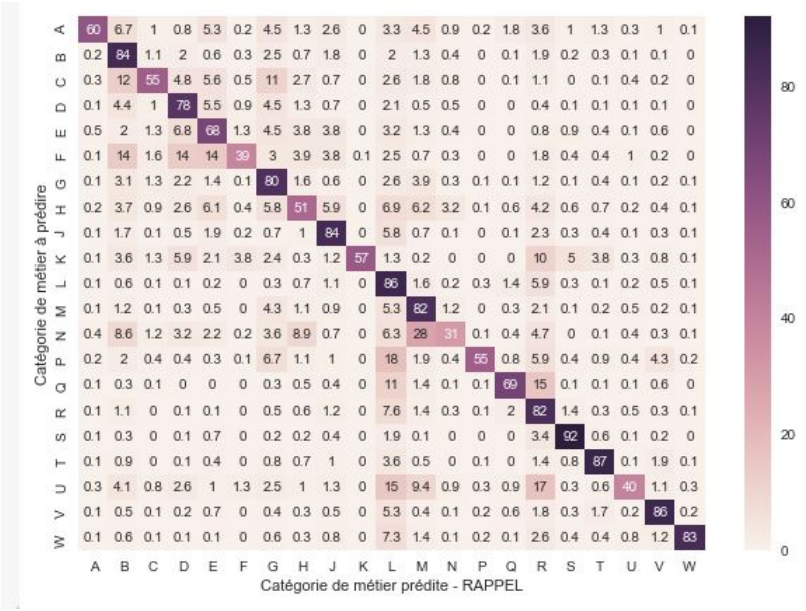

FIGURE 3 – Matrice de confusion normalisée pour le calcul du rappel

ceux de l'hôtellerie et de la restauration (catégorie S), les résultats obtenus sont très bons, ce qui peut s'expliquer par la présence de mots-clés très spécifiques (comme par exemple les mots « serveur » ou « restauration ») dans les offres d'emploi correspondant à ces métiers.

3 Discussion et travaux ultérieurs

On discute dans cette section des pistes potentielles pour améliorer les résultats présentés dans ce papier, ainsi que des autres difficultés soulevées par l'exploitation et l'analyse des offres d'emploi en ligne.

3.1 Autres pistes pour la codification du métier

D'abord, il y a plusieurs pistes pour améliorer les résultats pour la codification du métier recherché selon les offres d'emploi. Il apparaît en particulier opportun de réfléchir à la façon de sélectionner les variables utilisées pour la prédiction si on adopte des approches de type *machine learning*. En particulier, choisir des lemmes (mots) à utiliser comme prédicteur qui soient spécifiques de la variable à prédire (le métier recherché dans cette étude) permet d'améliorer les résultats, comme vu précédemment. Par ailleurs, on ne mobilise ici que le vocabulaire du descriptif de l'offre, il pourrait être intéressant de mobiliser d'autres variables si ces dernières sont disponibles. En particulier, certains sites publiant des offres d'emploi permettent d'avoir une information sur le secteur d'activité de l'entreprise recruteuse (souvent selon une nomenclature *ad hoc* toutefois), cette variable pourrait être intéressante à mobiliser pour améliorer la performance prédictive des modèles.

Une autre voie d'amélioration potentielle réside en la méthode utilisée pour la constitution de l'échantillon d'apprentissage dans les modèles de *machine learning*. Ici, la constitution des données d'apprentissage est réalisée de manière aléatoire. Par conséquent, les catégories de métiers avec peu d'observations peuvent être mal prédites simplement car on dispose de trop peu d'observations pour entraîner un algorithme performant. On constate en effet dans cette expérimentation que les taux de précision et de rappel obtenus sont assez faibles pour les catégories de métiers où on dispose de relativement peu d'observations dans les données utilisées. Une piste alors potentielle pour améliorer les résultats est de surreprésenter ces catégories dans les données d'apprentissage.

Enfin, on peut améliorer les résultats en choisissant une version différente de la variable à prédire (le métier au sens de la nomenclature des familles professionnelles). Par exemple, une expérience conduite sur les modèles de régression logistique et de perceptron multi-couches a consisté à modifier la précision de la variable à prédire, et donc du nombre de modalités parmi lesquelles les observations sont réparties. Au lieu de considérer les 22 catégories de la nomenclature des FAP codées sur un unique caractère, on considère les niveaux plus détaillés de nomenclature, avec respectivement 87 et 225 catégories de métiers à prédire. D'une part, on constate une diminution significative des performances globales des modèles (de l'ordre de 10 points de pourcentage pour le score F1 global en considérant 87 catégories, et de l'ordre de 13 points avec 225 modalités). Cela est dû au fait qu'il y a plus de catégories possibles à prédire. Toutefois, on constate également d'autre part, si on évalue les résultats avec le niveau agrégé de la nomenclature Fap, un léger gain en ce qui concerne les performances prédictives du modèle. Ces résultats demandent à être consolidés, mais ils laissent supposer que l'emploi des catégories les moins agrégées lors de l'entraînement des modèles permet une discrimination plus fine des offres d'emploi par rapport

à l'utilisation des variables agrégées sur les métiers. Pour améliorer la précision des résultats obtenus, une piste d'amélioration pourrait alors consister à une modification de la fonction de perte utilisée lors de l'entraînement, de sorte à pénaliser plus fortement les erreurs de classification à un niveau détaillé demeurant des erreurs lorsqu'on considère un niveau plus agrégé de nomenclature.

3.2 Autres problématiques méthodologiques

L'exploitation de données sur les offres d'emploi en ligne pose également d'autres questions avant de pouvoir réaliser des analyses statistiques. En particulier, nous nous sommes concentrés dans ce papier uniquement sur le codage du métier recherché, qui peut être une variable d'intérêt fort utile pour étudier les appariements entre offre et demande de travail. Il serait également intéressant de disposer d'informations complémentaires sur les offres d'emploi en ligne, notamment en ce qui concerne les établissements recruteurs (ou *a minima* des informations agrégés sur ces derniers comme leur secteur d'activité). On peut toutefois noter que les méthodes proposées dans ce papier pour codifier le métier peuvent être adaptées pour traiter ou contribuer à traiter l'identification d'autres informations.

Par ailleurs, la structuration des données sur les offres d'emploi en ligne selon les compétences recherchées par les recruteurs pourrait permettre d'étudier les désajustements sur le marché du travail en termes de compétences disponibles et recherchées (*skills mismatch*. Des travaux coordonnés par le Cedefop (Centre Européen de Développement de la Formation Professionnelle) sont notamment en cours pour extraire l'information sur les compétences contenue dans les offres d'emploi disponibles en ligne[4]. Les compétences seront alors codées selon la classification européenne ESCO, qui est une nomenclature permettant d'étudier à la fois les métiers et les compétences demandées.

Ensuite, dans l'optique de l'élaboration de statistiques représentatives, si on veut utiliser les offres d'emploi publiées en ligne pour compléter ou remplacer des dispositifs existants mobilisant des sources d'enquête ou des données administratives, il faut être capable de s'assurer de la stabilité du champ statistique considéré. Cela pose en particulier trois types de questions :

- D'abord, dans le cas de l'élaboration de statistique récurrentes, il faut être capable d'assurer un accès pérenne aux données, qui ne soit pas dépendant des évolutions techniques des sites ou des changements législatifs. Par ailleurs, les évolutions du marché de l'offre d'emploi en ligne peuvent conduire à des modifications dans la structure des sites publiant des offres d'emploi en ligne, notamment suite à l'émergence de nouveaux acteurs[7]. Comment s'assurer de la stabilité du champ couvert ?
- Une autre question non évoquée dans ce papier ici est celle de la déduplication des offres d'emploi publiées sur plusieurs sites internet, ou à la fois collectées par Pôle emploi et par ailleurs publiées sur le Net. IL est nécessaire de supprimer les doublons, ce qui peut représenter une tâche ardue car :
  - Deux offres apparemment très proches peuvent parfois correspondre à des postes différents, par exemple dans le cas où elles sont mises en ligne par une agence d'intérim qui veut limiter la charge de travail rédactionnelle.
  - Deux offres pour un même poste peuvent parfois être légèrement différentes selon les sites où elles sont postées si les personnes postant l'offre ont apporté des modifications au texte. Par ailleurs, certaines variables collectées par les sites dans certains cas (expérience requise par exemple) peuvent ne pas être cohérentes entre les différents sites, notamment si les nomenclatures adoptées sont différentes (par exemple, l'expérience requise pourra être « 0 à 5 ans d'expérience » pour un site et « 3 à 10 ans » pour un autre).

— Enfin, l'analyse de la demande de travail par le prisme des offres d'emploi publiées en ligne suppose que ces dernières sont encore pertinentes et n'ont pas été pourvues. La mobilisation des offres d'emploi collectées par Pôle emploi permet d'avoir (de manière imparfaite) cette information. En revanche, il est plus complexe de savoir si une offre publiée en ligne a été pourvue, notamment dans le cas où la publication de l'offre représente un service gratuit. La question de l'étude la pertinence des offres disponibles en ligne doit donc être soulevée.

En conclusion, il apparaît que les offres d'emploi en ligne peuvent représenter une source d'informations très intéressante pour l'étude des déterminants de la demande de travail. L'exploitation de telles données non structurées pose cependant plusieurs problèmes méthodologiques, notamment pour récupérer ces données et pour structurer l'information qu'elles contiennent. On a montré ici que, suite à la mobilisation de méthodes fondées sur des techniques de *matching* ou de *machine learning*, on pouvait coder l'information récupérée dans les données textuelles sur les offres d'emploi, notamment en ce qui concerne le métier recherché. Les résultats obtenus sur la performance des algorithmes de *machine learning* pour la codification du métier recherché sont encourageants. Au final, on peut toutefois noter que l'évaluation de la performance des algorithmes mobilisés pour structurer l'information n'est possible qu'à condition de disposer de données pouvant être utilisées comme jeu de test. Cela pose donc la question du **benchmark** : quelles données de référence utiliser pour structurer les informations récupérées en ligne ? Comment s'assurer de la qualité de la structuration des données si jamais on cherche à remplacer (plutôt que compléter) les sources déjà existantes, qui ont ici été utilisées pour constituer les données d'apprentissage et de validation ?

Références

[1] Bergeat M. (2017), « Les tensions sur le marché du travail au 2e trimestre 2017 », *Dares Indicateurs*, n° 056.

[2] Bergeat M., Rémy V. (2017), « Comment les employeurs recrutent-ils leurs salariés ? », *Dares Analyses*, n° 064.

[3] Breiman L. (2001), "Random Forests", *Machine Learning*, vol. 45, p. 5-32.

[4] Cedefop (2016), "Using labour market information", *Cedefop series of guides on skills anticipation and matching*, Centre européen pour le développement de la formation professionnelle, n° 1.

[5] ESSNet Big Data (2017, 2018), "Final technical reports", *Webscraping job vacancies work package*, disponibles ici.

[6] Fondeur Y. (2016), « Les offres d'emploi sur Internet : vers la 'transparence' du marché du travail ? », *Connaissance de l'emploi*, Centre d'études de l'emploi, n° 132.

[7] Fondeur Y. (2017), « Google et le marché numérique du travail », *Connaissance de l'emploi*, Centre d'études de l'emploi, n° 136.

[8] Murphy, K. P. (2006). "Naive bayes classifiers". *University of British Columbia*, vol. 18.

[9] Rumelhard D., Hinton, G., William R (1986). "Learning representations by back-propagating errors". *Nature*, vol. 323, n° 6088, p. 533.

[10] Vroylandt T. (2018), « Les offres d'emploi diffusées par Pôle emploi au 4e trimestre 2017 », *Statistiques, études et évaluation*, Pôle emploi, n° 18.006.