RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Dares
Déchiffrer le monde du travail
pour **éclairer** le débat public

• DOCUMENT D'ÉTUDES •

MARS 2022
N° 258

Les offres d'emploi en ligne, nouvelle source de données sur le marché du travail : illustration sur l'année 2019

Claire de Maricourt
Moustapha Niang
Dares

Les offres d'emploi en ligne, nouvelle source de données sur le marché du travail

Illustration sur l'année 2019

Claire de Maricourt et Moustapha Niang

Résumé

En quelques années, Internet est devenu une nouvelle source d'information sur le marché du travail. Selon l'enquête Offre d'emploi et recrutement (Ofer) de la Dares, 95 % des annonces d'offres d'emploi ont fait l'objet d'une publication sur Internet en 2016 contre 53 % en 2005. Forte de ce constat, la Dares a décidé de collecter les offres d'emploi en ligne publiées sur une quinzaine de sites pour en faire une nouvelle base de données sur les offres d'emploi : Jocas (*Job offers collection and analysis system).* Différents outils sont utilisés pour construire cette nouvelle base de données : *webscraping*, algorithme de classification automatique de texte, déduplication.

Sur l'année 2019, la base Jocas peut être comparée aux sources usuelles de la statistique publique sur l'offre d'emploi, qu'il s'agisse de sources administratives, comme les offres diffusées par Pôle emploi et les Déclarations préalables à l'embauche (DPAE) des Urssaf, ou bien des données issues d'enquête telles que celle sur les Besoins en main-d'œuvre (BMO) de Pôle emploi, l'enquête Emploi de l'Insee, l'enquête Activité et conditions d'emploi de la main-d'œuvre (Acemo) de la Dares. Il en ressort que les métiers sont inégalement couverts par Jocas. Les domaines professionnels avec une forte proportion de cadres ou effectuant beaucoup de recrutements en ligne ont tendance à être surreprésentés. Au contraire, ceux comptabilisant beaucoup de recrutements multiples ou mobilisant des canaux de recrutement informels sont plutôt sous-représentés.

La base Jocas fait déjà l'objet d'exploitations statistiques. Les données d'offres en ligne ont notamment été intégrées au calcul des tensions sur le marché du travail. Elles ont également été utilisées pour la production du tableau de suivi de la situation du marché du travail en 2020-2021 lors de la crise du Covid-19.

Mots-clés : offres d'emploi en ligne, *webscraping*, Big data, classification de texte, NLP, représentativité, indice de dissimilarité de Duncan
Codes JEL : J23, J63, C43, C81

Remerciements

Nous remercions vivement et en premier lieu **Bertrand Lhommeau** qui a supervisé et coordonné cette étude. Nos remerciements s'adressent également à nos collègues successifs du Département Analyse des métiers et emploi des travailleurs handicapés (Dares/Dameth) qui ont travaillé - avec ou avant nous - sur le développement du projet : **Maxime Bergeat, Paul Andrey** et **Alexis Eidelman**. Enfin, nous remercions **Yannick Fondeur** du Cnam (Lise/CEET) pour son expertise et ses commentaires avisés.

Table des matières

Introduction

Ces deux dernières décennies ont vu s'accroître l'utilisation des outils numériques dans les sphères personnelle et professionnelle (COUSTEAUX, 2019). Le marché du travail ne fait pas exception et Internet est désormais un canal de recrutement fréquemment utilisé par les employeurs (BERGEAT et *al.*, 2018). La part des offres d'emploi ayant fait l'objet d'une publication en ligne a notamment fortement augmenté : d'après l'enquête sur les offres d'emploi et les recrutements (Ofer) menée par la Dares, cette part est passée en France de 53 % en 2005 (BESSY et MARCHAL, 2006) à 95 % en 2016 (BERGEAT et REMY, 2017). Parallèlement à l'expansion du marché des offres d'emploi en ligne, de nombreux outils informatiques pour la collecte et l'exploitation de données massives provenant d'Internet sont apparus. La plupart des langages de programmation « grand public » intègrent des fonctionnalités dédiées à la collecte de données en ligne et facilitant leur traitement. Ce nouveau gisement de données immédiatement accessibles suscite un fort intérêt auprès d'acteurs non institutionnels : certains sites diffuseurs d'offres en ligne se sont mis à publier des statistiques sur leurs données et des acteurs privés proposant des analyses du marché du travail fondées sur ce type de données en ligne ont émergé.

Bien que les acteurs publics disposent déjà de nombreuses sources de données sur le marché du travail (données administratives ou données d'enquêtes), ce changement de contexte incite les instituts statistiques à se positionner par rapport aux données Internet et à envisager de nouvelles méthodes d'analyse du marché du travail. Ainsi, Eurostat lançait en 2016 le premier « ESSnet Big data » afin de coordonner les travaux exploratoires des services statistiques européens sur le sujet. S'inscrivant dans ce projet, la Dares collecte quotidiennement depuis fin 2018 des offres d'emploi publiées en ligne sur un échantillon de sites de recrutement. Cette collecte massive a permis la création d'une nouvelle base de données contenant environ 5 millions d'offres d'emploi pour l'année 2019 : Jocas (*Job offers collection and analysis system*).

L'objectif de ce document d'étude est de répondre aux principales questions statistiques soulevées par l'émergence des données d'offres en ligne : comment produire une base statistique à partir de données semi-structurées collectées en ligne ? Quels sont les apports mais aussi les limites de la source Internet par rapport aux sources traditionnelles sur l'emploi ? Quels nouveaux indicateurs proposer à partir de ces données ?

La première partie du document présente le contexte dans lequel s'inscrivent ces travaux. Au-delà des définitions, elle positionne les offres d'emploi en ligne au sein du marché du travail et fait un état de l'art des recherches sur cet objet. La deuxième partie détaille la méthodologie utilisée à la Dares pour collecter les offres d'emploi en ligne (*webscraping*) et décrit les étapes de la création d'une base statistique à partir de ces données, notamment la codification automatique des offres par métier et la déduplication. Enfin, la troisième partie met en perspective la source Internet par rapport aux sources traditionnelles sur le marché du travail et s'attache à expliquer leurs points de divergence. Cette dernière partie ouvre aussi la voie aux usages possibles de ces données et propose des pistes d'utilisation de la base Jocas.

I. Émergence d'une source en ligne sur le marché du travail

1. Offre(s) d'emploi et offres en ligne

L'offre d'emploi (ou demande de travail) correspond, à un instant donné, aux besoins de main-d'œuvre exprimés par les employeurs afin de pouvoir réaliser leur production de biens et services. En pratique, ce concept se mesure aujourd'hui grâce à la notion d'emplois vacants dont la direction générale de la Commission européenne chargée de l'information statistique à l'échelle communautaire (Eurostat) propose la définition suivante[1] :

> « Une vacance d'emploi désigne tout poste rémunéré nouvellement créé, inoccupé ou sur le point de devenir vacant,
>
> > (a) pour lequel l'employeur entreprend activement de chercher, en dehors de l'entreprise concernée, un candidat apte et est prêt à entreprendre des démarches supplémentaires,
> >
> > (b) et qu'il a l'intention de pourvoir immédiatement ou dans un délai déterminé (trois mois). »

Le recrutement souhaité peut correspondre à un contrat à durée indéterminée (CDI), un contrat à durée déterminée (CDD), ou à un emploi saisonnier, même de courte durée. Les emplois vacants correspondent donc à un stock d'emplois, au moment d'une enquête, pour lesquels les employeurs cherchent un candidat.

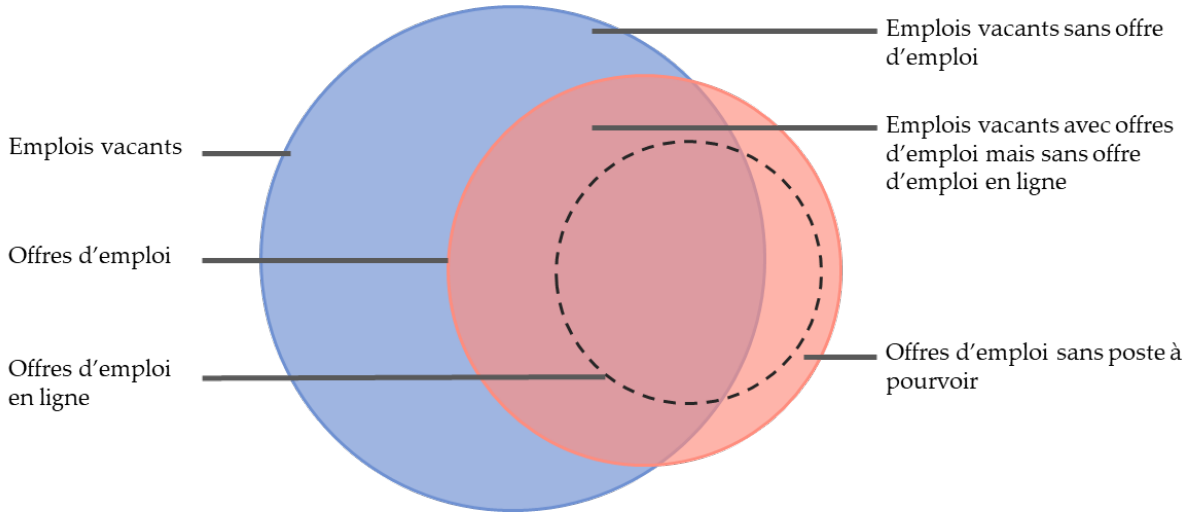
Pour trouver le candidat adéquat, les employeurs peuvent activer différents canaux de recrutement (BERGEAT et REMY, 2017) notamment la publication d'une offre d'emploi, une annonce publique de leur besoin d'emploi. Les offres d'emploi doivent respecter certains critères légaux[2] et contenir les informations indispensables à la description d'un emploi[3] : une date de publication, un intitulé de métier, un lieu de travail, des informations sur l'employeur, sur l'expérience requise,... Parmi ces offres d'emploi, certaines sont diffusées sur Internet : ce sont les offres d'emploi en ligne. Cependant, toutes les vacances d'emploi ne donnent pas lieu à des offres d'emploi – en ligne ou non. De même, comme cela sera expliqué par la suite, toutes les offres d'emploi diffusées ne correspondent pas nécessairement à des emplois vacants (Figure 1). Emplois vacants, offres d'emploi et offres d'emploi en ligne sont donc étroitement liés et sont autant de *proxies* de la demande de travail.

[1] https://ec.europa.eu/eurostat/fr/web/labour-market/job-vacancies

[2] https://travail-emploi.gouv.fr/droit-du-travail/la-vie-du-contrat-de-travail/article/offre-d-emploi-et-embauche-les-droits-du-candidat

[3] Rédiger une offre d'emploi : quelques règles de base |Pôle emploi (pole-emploi.fr)

FIGURE 1 - Emplois vacants, offres d'emploi et offres d'emploi en ligne



Note : ce graphique est une représentation schématique, les proportions n'y sont pas respectées.
Lecture : un emploi vacant peut donner lieu à une publication d'offre – en ligne ou non. Toutes les offres d'emploi ne correspondent pas à des postes à pourvoir.

2. Les sources usuelles de la statistique publique

En France, il existe plusieurs sources de données sur les offres d'emploi et les emplois vacants. Ces sources peuvent être :

- des enquêtes statistiques telles que l'enquête Besoins en main-d'œuvre (BMO) de Pôle emploi, l'enquête Activité et conditions d'emploi de la main-d'œuvre (Acemo) de la Dares et l'enquête Emploi en continu (EEC) menée par l'Insee et qui a vocation à observer le marché du travail de manière structurelle et conjoncturelle (notamment en fournissant la mesure des concepts d'activité, de chômage, d'emploi et d'inactivité tels qu'ils sont définis par le Bureau international du travail (BIT)) (Tableau 1).
- des extraits de fichiers administratifs tels que la Statistique du marché du travail (STMT) issue des offres et demandes d'emploi déposées à Pôle emploi qui sont produites par Pôle emploi et la Dares, les Déclarations préalables À l'embauche (DPAE) gérées par la Mutualité sociale agricole (MSA) et les Urssaf (Union de recouvrement des cotisations de sécurité sociale et d'allocations familiales) et les Mouvements de main-œuvre (MMO) issues des Déclarations sociales nominatives (DSN) et élaborés par la Dares.

Chacune de ces sources vise cependant un champ différent. La STMT couvre l'ensemble des offres d'emploi collectées par Pôle emploi ; les DPAE concernent les déclarations d'intentions d'embauche ; l'EEC permet d'estimer le nombre d'emploi ayant débuté depuis moins d'un an ; Acemo mesure les emplois vacants ; BMO fournit une estimation du nombre de projets de recrutement ; MMO permet de recenser le nombre d'embauches.

La STMT collecte les offres déposées par les recruteurs à Pôle emploi. Elle permet d'obtenir des données individuelles sur ces offres d'emploi et est disponible sur une base mensuelle. Elle est déclinable par métier, secteur d'activité et localisation géographique (région, département et zone d'emploi). Elle ne couvre cependant que partiellement l'ensemble du marché du travail, de l'ordre

d'un tiers seulement de l'ensemble des recrutements (COE, 2013 ; BERGEAT et *al.*, 2018). Ce défaut de couverture varie selon les métiers et/ou les secteurs étudiés ainsi que les types de contrats ; il évolue au cours du temps et a tendance à augmenter avec la multiplication du nombre d'acteurs dans le secteur de la diffusion d'offres d'emploi en ligne (COE, 2013).

Les DPAE sont des données administratives permettant de mesurer les intentions d'embauches en France hors Mayotte, sur l'ensemble des activités concurrentielles et le secteur public sur le champ des contrats de droit privé. Ces données peuvent être déclinées selon le type de contrat proposé dans l'intention d'embauche et le secteur d'activité de l'établissement recruteur. Cependant, les données ne sont pas déclinables directement par métier et seules les intentions d'embauches sont observées : la source ne fournit aucune information sur l'aboutissement de ces intentions et ne couvre pas les emplois vacants ou les offres d'emploi qui n'ont pas donné lieu à une embauche ou une intention d'embauche.

L'EEC s'inscrit dans le cadre des enquêtes « Forces de travail » défini au niveau européen ("Labour Force Survey")[4] et vise à observer le marché du travail de manière structurelle et conjoncturelle. C'est la seule source fournissant une mesure des concepts d'activité, de chômage, d'emploi et d'inactivité tels qu'ils sont définis par le Bureau international du travail (BIT). Elle date de 1950 et est devenue continue et trimestrielle à partir de 2003. Au fil des décennies, l'enquête a intégré des changements de natures diverses : des changements de questionnaires et de concepts (notamment pour se conformer aux orientations du BIT ou d'Eurostat), mais aussi des évolutions méthodologiques (sur l'échantillonnage ou le traitement de la non-réponse par exemple) ou encore techniques (sur les modes de collecte, l'informatisation du traitement des données…). Cette source permet de mesurer le nombre d'emplois ayant débuté depuis moins d'un an (sur un trimestre ou sur l'année), mais, comme la base des DPAE, ne permet pas d'estimer les projets de recrutement qui ne se concrétisent pas.

L'enquête Acemo permet d'approcher la vacance d'emploi (selon la définition d'Eurostat, cf. I.1) sur les salariés du secteur privé hors agriculture, particuliers employeurs et activités extraterritoriales en France (hors Mayotte). Elle est réalisée sur une base trimestrielle avec un échantillon conséquent (35 000 établissements répondants) permettant une déclinaison par secteur d'activité. En complément à cette enquête trimestrielle portant sur les établissements d'au moins 10 salariés, une enquête est menée chaque année auprès d'un échantillon d'environ 60 000 entreprises de moins de 10 salariés.

Cependant, la définition européenne retenue pour l'indicateur du taux d'emplois vacants présente deux limites :

- il n'y a pas d'indication temporelle concernant l'emploi « sur le point de se libérer » et les « démarches actives » entreprises ne sont pas précisées ;
- la mesure ne permet pas d'appréhender un nombre d'emplois « durablement vacants », la date de début de recherche n'étant pas connue. Cette notion d'emploi vacant ne renseigne donc pas sur les difficultés de recrutement.

En outre, la vacance d'emploi n'est pas déclinable par métier (COE, 2013). Enfin, l'enquête Acemo comporte plusieurs ruptures de séries sur la période récente et ne couvre pas le secteur public.

[4] https://ec.europa.eu/eurostat/web/lfs

L'enquête BMO est réalisée annuellement par Pôle emploi sur un échantillon de taille importante d'établissements du secteur privé (400 000) et permet de mesurer le nombre de projets de recrutement en métropole et dans les Dom selon le métier, le secteur d'activité et la localisation géographique des établissements. Elle permet aussi de distinguer les projets de recrutement saisonniers et ceux anticipés comme difficiles par les recruteurs. L'enquête comporte depuis 2011 un volet complémentaire auprès de 20 000 des établissements précédemment interrogés sur les motifs de recrutement ou d'abandon de projets de recrutement ainsi que sur la nature des difficultés anticipées. Les limites principales de la source concernent son champ, qui ne couvre pas l'ensemble de la fonction publique, ainsi que son taux de réponse, plus faible notamment pour les établissements de petite taille et variable suivant les secteurs d'activité.

Le concept de recrutement anticipé mesuré avec l'enquête BMO a plusieurs limites :

- Il peut être compliqué pour les établissements d'anticiper avec précision les recrutements, sur un horizon de 12 à 16 mois compte tenu du calendrier de l'enquête. En particulier, le recruteur ne peut pas estimer les éventuels départs (qui seront donc à remplacer durant l'année) des personnes qui n'ont pas été encore recrutées : un besoin en main-d'œuvre peut se traduire par plusieurs recrutements sur une même année.
- Dans le questionnaire, on ne distingue pas les besoins en main-d'œuvre anticipés en fonction du type de contrats à pourvoir, à l'exception des contrats saisonniers.

Les MMO sont des données administratives (issues des DSN depuis 2018) permettant de recenser les embauches (ou entrées) et les fins de contrats (ou sorties) par nature de contrats (CDI ou CDD mais hors missions d'intérim) sur la base des déclarations des établissements employeurs de France métropolitaine des secteurs privés hors agriculture et particuliers employeurs. Ces données peuvent être déclinées selon le type de contrat et le secteur d'activité des établissements. Cependant, les données ne sont à ce stade pas déclinées par métier. De plus, par essence, cette source renseigne sur les embauches effectives ; elle ne couvre pas les emplois vacants ou les offres d'emploi qui n'ont pas donné lieu à un recrutement.

TABLEAU 1 - Sources de la statistique publique sur les offres d'emploi

Source	Type de source	Champ et unité	Nombre d'observations sur une année	Profondeur et granularité temporelle	Déclinaison par secteur / par métier	Concepts de demande de travail	Principales limites
Acemo	Enquête statistique	Etablissements du secteur privé hors agriculture, particuliers employeurs et activité extra-territoriales (France hors Mayotte)	~200 000	2003 Trimestrielle (annuelle pour les établissements de moins de 10 salariés)	Oui / Non	Emplois vacants	- Un concept d'emploi vacant à interpréter avec prudence - Deux ruptures de séries récentes pour les emplois vacants
BMO	Enquête statistique	Etablissements du secteur privé (France hors Mayotte)	~400 000	2002 Annuelle	Oui / Oui	Projets de recrutement	- Incertitude de la mesure liée à la difficulté d'anticipation des projets de recrutement - Taux de réponse d'environ 25 %
STMT	Données administratives	Champ des offres collectées par Pôle emploi et des demandeurs inscrits à Pôle emploi (France hors Mayotte)	~8 millions (offres d'emploi) ~90 millions (DEFM)	1998 Mensuelle	Oui / Oui	Offres d'emploi collectées par Pôle emploi	- Restriction du champ en fonction de la couverture de Pôle emploi
DPAE	Données administratives	Activités concurrentielles et contrats de droit privé du secteur public (France hors Mayotte)	~40 millions	2000 Mensuelle	Oui / Non	Intentions d'embauche	- Pas d'information sur les recrutements qui n'aboutissent pas
EEC	Enquête statistique	Individus de 15 ans ou plus vivant en logement ordinaire. (France hors Mayotte)	~450000	Annuelle depuis 1950 Trimestrielle à partir de 2003	Oui / Oui	Emplois de moins d'un an	- Pas d'information sur les recrutements qui n'aboutissent pas - Faible taille de l'échantillon pour une analyse fine par métier
MMO	Données administratives	Secteur privé hors agriculture, intérim et particuliers employeurs (France métropolitaine)	~25 millions	2007 Trimestrielle	Oui / Non	Embauches	Pas d'information sur les recrutements qui n'aboutissent pas

3. Le marché numérique du travail comme nouvelle source de données

a. Une source multiforme

Loin de constituer une source homogène, le marché numérique du travail regroupe de nombreux sites dont les modèles de diffusion d'offres d'emploi varient fortement. La littérature fournit une typologie détaillée des principaux types d'acteurs de la diffusion d'offres sur Internet (FONDEUR et LHERMITTE, 2006 ; COE, 2015 ; FONDEUR, 2016). Aujourd'hui, les offres d'emploi sur Internet sont ainsi principalement accessibles sur :

- Les sites d'emploi « propriétaires », comme les sites « carrière » des entreprises et les sites d'agences de recrutement ou d'intérim, qui diffusent leurs propres offres.

- Les *job boards*, qui sont des intermédiaires directs entre candidats et recruteurs. Ils diffusent auprès des candidats les offres d'emploi publiées par les recruteurs sur le site[5].
- Les agrégateurs, qui sont des moteurs de recherche d'offres d'emploi. Ils indexent des *job boards* ou des sites « propriétaires » afin de proposer au candidat un large panel d'offres d'emploi. Si un candidat veut postuler, il est redirigé vers le site d'origine de l'offre.
- Les réseaux sociaux exclusivement dédiés à l'emploi comme LinkedIn et certains réseaux sociaux généralistes (par exemple Facebook, qui permet aux entreprises de diffuser des offres sur la plateforme à l'aide de l'outil « Facebook jobs »).
- Les médias en ligne spécialisés proposant une rubrique pour le partage d'annonces (le site de « L'étudiant » dispose par exemple d'une page dédiée aux offres d'emploi).

Les frontières entre les différents types de sites d'emploi sont toutefois assez poreuses. D'une part, il semble qu'aujourd'hui la plupart des agrégateurs ne se contentent plus d'indexer des contenus existants mais permettent aussi aux recruteurs de déposer des offres sur leur site et aux candidats d'y postuler. D'autre part, certains *job boards* traditionnels diffusent désormais des offres provenant de sites tiers. C'est notamment le cas de Pôle emploi, qui, en plus de diffuser l'ensemble des offres déposées à Pôle emploi sur le site pole-emploi.fr, publie depuis 2013 sur son site des offres provenant de sites d'emploi partenaires (BERGEAT et *al.*, 2018).

Au-delà des différences de collecte et de diffusion des offres, le type et la qualité des informations contenues dans les offres peuvent largement varier d'un site d'emploi à un autre. Certains sont spécialisés dans un secteur (par exemple, l'agriculture), certains types de métiers (comme les métiers de la garde d'enfants), un lieu ou une catégorie d'emploi (les plateformes de micro-travail sont notamment spécialisées dans l'emploi non salarié). Par ailleurs, les sites sont payants ou gratuits, publics ou privés. La durée maximale d'apparition d'une offre sur un site varie également.

Bien que le paysage des sites d'offres d'emploi et les pratiques des employeurs évoluent rapidement[6], l'analyse de l'utilisation des sites Internet par les recruteurs dans l'enquête Ofer de 2016 (Dares) fournit un cadrage du marché des types de sites d'offres d'emploi[7]. Fin 2015, Pôle emploi était de loin le site le plus mobilisé : pour 62 % des recrutements[8] ayant fait l'objet d'une annonce en ligne, une offre d'emploi a été diffusée sur le site de Pôle emploi (hors sites partenaires, Figure 2). Le site de l'entreprise était lui aussi largement utilisé par les recruteurs (40 %) et devançait les *job boards* généralistes (14 %) et spécialisés (18 %). Enfin, les recruteurs ont assez peu diffusé d'annonce en ligne sur les réseaux sociaux en 2016 (2 %).

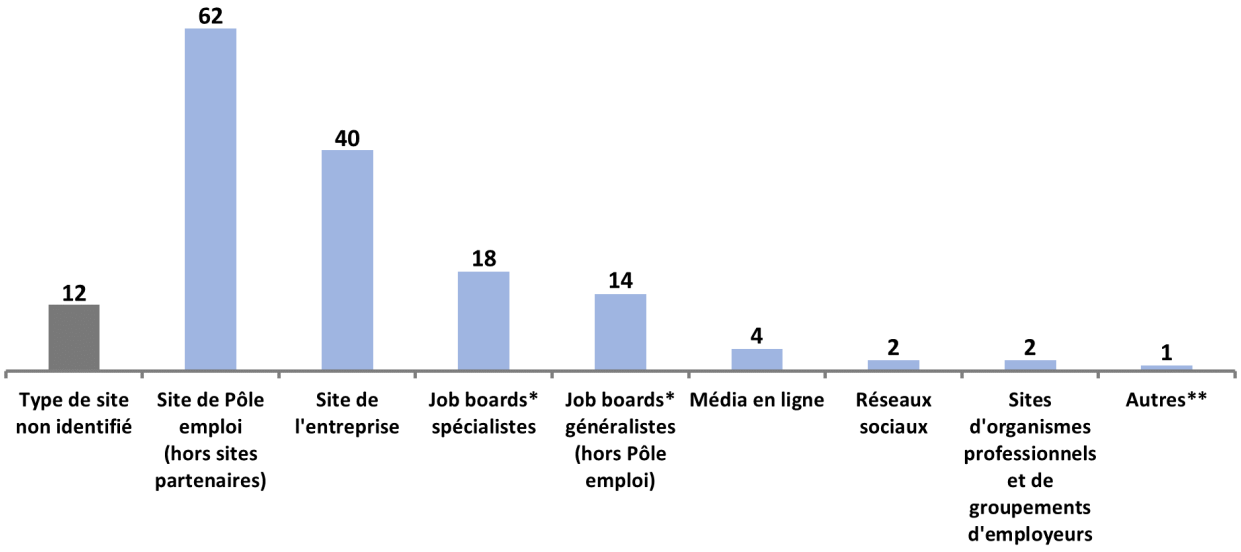
[5] Nous incluons ici dans les *job boards* les sites d'annonces classées (ou petites annonces) tels que définis par FONDEUR (2016).

[6] À titre indicatif, sur les 14 sites scrapés en 2019, 2 avaient disparu au 1er janvier 2021.

[7] Hormis les agrégateurs, qui par définition ne sont pas utilisés directement par les recruteurs (sauf s'ils ont une activité de *job board*).

[8] L'enquête Ofer est restreinte aux recrutements en CDI et en CDD de plus d'un mois. Les offres d'emploi n'ayant pas abouti à un recrutement ou les offres pour des contrats courts sont ici exclues.

FIGURE 2 - Taux de recours aux différents types de sites parmi les recrutements avec diffusion d'une annonce en ligne



* y compris les agrégateurs ayant une activité de job board
** établissements d'enseignement et de formation, missions locales, cabinets de recrutement ou d'intérim

Note : hormis pour le site de Pôle emploi et le site de l'entreprise, la typologie a été établie de manière *ad hoc* à partir des *verbatims* de noms de sites renseignés par les recruteurs dans l'enquête Ofer. L'ensemble des offres de Pôle emploi donnent lieu à une publication d'offre en ligne.

Lecture : une offre d'emploi est diffusée sur le site de l'entreprise pour 40 % des recrutements ayant fait l'objet d'une annonce en ligne. Pour 12 % des recrutements passant en ligne, le ou les types de sites utilisés par le recruteur n'ont pas été identifiés.

Champ : ensemble des recrutements en CDI ou en CDD de plus d'un mois entre septembre et novembre 2015, au sein des établissements d'au moins un salarié du secteur concurrentiel ayant fait l'objet d'une annonce sur Internet ; France.

Source : Dares, enquête Ofer 2016.

b. Une source à fort potentiel, mais pas sans limites statistiques

La relative facilité d'acquisition des offres en ligne et les informations nouvelles qu'elles délivrent contribuent à l'amélioration de la connaissance du marché du travail. Toutefois, les apports de cette source ne doivent pas en occulter les limites statistiques, et notamment les interrogations liées à sa représentativité.

L'exploitation des millions d'offres d'emploi publiées en ligne chaque année peut améliorer la compréhension du marché du travail. En effet, des informations qui n'étaient pas captées jusqu'à maintenant (comme la description complète de l'offre ou les date et durée de mise en ligne) sont désormais accessibles. Contrairement aux sources usuelles, l'accès à ces données se fait sans intermédiaire et théoriquement de manière instantanée. L'analyse des offres sur Internet permet donc un suivi haute fréquence de l'offre d'emploi, à moindre coût.

En revanche, contrairement aux enquêtes et aux sources administratives, ces données reflètent des effets liés à des stratégies de recrutement et qui peuvent biaiser les analyses quantitatives. Par exemple, pour maximiser leurs chances de trouver un candidat adéquat, les recruteurs choisissent souvent de diffuser une même offre sur plusieurs sites – notamment à l'aide de prologiciels de « multidiffusion » d'offres en ligne (FONDEUR et LHERMITTE, 2013). À l'inverse,

certains recruteurs laissent volontairement en ligne des offres fictives qui ne servent qu'à collecter des CV en anticipation d'éventuels besoins futurs.

L'utilisation de cette nouvelle source soulève également des questions statistiques majeures. Bien que l'enquête Ofer renseigne sur la part de marché de certains types de sites, il est difficile d'évaluer le champ couvert par un site ou un panel de sites. De plus, la source Internet est instable et mouvante : les émetteurs d'offres changent et se transforment rapidement et le champ couvert n'est donc *a priori* pas constant. La représentativité des offres en ligne parmi l'ensemble des recrutements questionne également. En effet, les 49 % de recrutements faisant l'objet d'une annonce sur Internet (BERGEAT et *al.*, 2018) ne sont pas représentatifs de l'ensemble des recrutements : le recours au canal Internet varie fortement selon les métiers. Près de 66 % des recrutements dans le domaine de l'informatique et des télécommunications font l'objet d'une annonce en ligne (Tableau 2), mais seuls 29 % des recrutements dans le bâtiment passent par ce canal. Néanmoins, l'écart (mesuré par l'indice de dissimilarité de Duncan) entre la répartition par domaine professionnel de l'ensemble des recrutements et celle des recrutements avec diffusion d'une annonce en ligne est finalement réduit (6,5 %) (Tableau 2). En 2016, la structure par domaine professionnel de l'ensemble des recrutements en ligne était même plus proche de l'ensemble des recrutements que de celle des recrutements passant par Pôle emploi (hors sites partenaires de Pôle emploi : dissimilarité avec l'ensemble des recrutements de 7,9 %). Enfin, bien que certains métiers aient plus recours aux offres d'emploi en ligne que d'autres selon l'enquête Ofer, ils peuvent être peu présents dans une source de données en ligne. Par exemple, le domaine professionnel des ingénieurs et cadres de l'industrie et celui de l'informatique et des télécommunications sont surreprésentés dans l'ensemble des recrutements en ligne alors qu'ils sont sous-représentés sur le seul site de Pôle emploi.

TABLEAU 2 - Part de recrutements en ligne et contribution des domaines professionnels à la dissimilarité de Duncan avec l'ensemble des recrutements

Domaines professionnels (FAP 22)	Part de recrutements en ligne	Contribution à la dissimilarité de Duncan avec l'ensemble des recrutements	
		Recrutements avec diffusion d'une annonce en ligne	Recrutements avec diffusion d'une annonce sur le site de Pôle emploi
Bâtiment, travaux publics	29	1,9	1,7
Électricité, électronique	35	0	0
Mécanique, travail des métaux	30	0,5	0,4
Industries de process	40	0,1	0,1
Matériaux souples, bois, industries graphiques	54	0	0,1
Maintenance	53	0,2	0,5
Ingénieurs et cadres de l'industrie	54	0,1	0,3
Transports, logistique et tourisme	51	0,2	1
Artisanat	41	0	0,1
Gestion, administration des entreprises	54	0,6	0,2
Informatique et télécommunications	66	0,4	0,5
Études et recherche	54	0	0,2
Banque et assurances	64	0,2	0
Commerce	51	0,5	0
Hôtellerie, restauration, alimentation	45	0,4	0,2
Services aux particuliers et aux collectivités	55	0,7	2,2
Communication, information, art et spectacle	48	0	0,5
Santé, action sociale, culturelle et sportive	51	0,2	0,1
Enseignement, formation	39	0,2	0,1
Total	**49**		
Dissimilarité de Duncan		**6,5**	**7,9**

Lecture : dans le domaine de l'informatique et des télécommunications, 66 % des recrutements font l'objet d'une diffusion d'annonce en ligne. Dans ce domaine, les offres en ligne comptent relativement plus d'offres que de recrutements (en jaune). Ce domaine est sous-représenté dans les offres déposées sur le site de Pôle emploi (en rouge). L'indicateur de Duncan indique que 6,5 % des recrutements avec diffusion d'une annonce en ligne (dont 1,9 % dans le bâtiment et les travaux publics) devraient changer de domaine professionnel pour que leur répartition s'aligne sur celle de l'ensemble des recrutements.

Champ : ensemble des nouveaux recrutements en CDI ou en CDD de plus d'un mois entre septembre et novembre 2015 des établissements d'au moins un salarié du secteur concurrentiel à l'exception des domaines professionnels agriculture, marine, pêche, administration publique, professions juridiques, armée et police et politique, religion ; France.

Source : Dares, enquête Ofer 2016.

4. Offres en ligne : un champ d'étude de plus en plus large

Depuis les premières études sur les données issues d'Internet, la mobilisation des données d'offres en ligne pour améliorer la connaissance de l'offre et de la demande de travail s'est faite de plusieurs manières. Les données d'offres d'emploi ont souvent été associées à la demande d'emploi, afin d'analyser le processus d'appariement sur le marché du travail. Ces analyses bifaces du marché du travail – qui se restreignent généralement à l'étude d'un site donné – permettent de mieux comprendre deux aspects du recrutement en ligne : le processus (DAVIS et SAMANIEGO DE LA PARRA, 2017) et la qualité de l'appariement *via* les phénomènes de *match* (BANFI et *al.*, 2018) et de *mismatch* sur le marché du travail (SINCLAIR et GIMBEL, 2020). Les offres d'emploi se sont aussi imposées comme des objets d'études en tant que tels. D'une part, des indicateurs, comme ceux sur la concentration géographique de l'offre d'emploi sur le marché du travail (AZAR et *al.*, 2020), ont vu le jour. D'autre part, des analyses du contenu textuel des offres ont émergé ce qui a notamment donné lieu à des publications sur les compétences recherchées par les employeurs (CEDEFOP, 2019).

Plus récemment, on assiste au développement de la collecte massive d'offres d'emploi en ligne auprès de multiples diffuseurs dans le but de couvrir l'ensemble des offres en ligne sur le marché du travail et d'estimer l'offre d'emploi globale. Des travaux en ce sens ont notamment été engagés par Eurostat depuis 2016[9]. Ce dernier usage, visant une couverture de la totalité du marché du travail d'un pays ou d'une communauté de pays est plus exigeant (KUREKOVA et *al.*, 2014). Ces travaux restent aujourd'hui le plus souvent exploratoires car ils posent des problèmes de représentativité (cf. ci-dessus) : ils concernent uniquement la partie du marché du travail couverte par le ou les site(s) diffuseur(s), dont la part est souvent difficile à estimer.

[9] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data#WPB_Online_job_vacancies

II. Produire une donnée de qualité statistique depuis une source Internet

1. Collecte des données en ligne : le choix du *scraping*

a. Choisir des sites sources

En France, les recruteurs ont accès à de nombreux sites d'offres d'emploi en ligne. Pour cette première phase expérimentale, le nombre de sites a été limité (15 maximum) afin de ne pas démultiplier les problèmes techniques et administratifs liés à la collecte.

Pour assurer une plus grande représentativité malgré leur nombre limité, les sites ont été choisis de manière à créer un ensemble couvrant tout le territoire, tous les types de métiers, de contrats et de qualifications. De plus, pour être retenus, les sites devaient diffuser des offres accessibles publiquement, sans avoir à se connecter à un compte utilisateur ou à payer par exemple. Pour assurer un volume minimum d'offres, des seuils ont été fixés : un stock journalier d'au moins 5 000 offres pour les sites spécifiques et 10 000 pour les sites généralistes. Enfin, les sites devaient être édités par des organismes dont le siège social est en France.

Après avoir expertisé plusieurs sites[10], un panel restreint mais diversifié a été élaboré. Il est composé de *job boards* français généralistes, d'un agrégateur et d'un ensemble de sites locaux couvrant toute la France[11]. La collecte automatisée a été lancée sur ces sites en septembre 2018. Depuis, d'autres sites ont été inclus : deux sites spécialisés sur les offres de cadre (en mars 2019 pour l'un et en octobre 2019 pour l'autre) et un site spécialisé dans les métiers du numérique (en octobre 2019). Ce panel de 14 sites n'est pas figé et est amené à évoluer.

b. Mode de collecte : le *scraping*, une possibilité parmi d'autres

Une fois les sources choisies, plusieurs possibilités existent pour collecter les offres. Trois méthodes de collecte automatisée d'offres d'emploi ont notamment été identifiées : le développement de partenariats avec des sites diffuseurs, la récupération de données *via* des API[12] existantes et enfin le *webscraping (*ou *scraping*).

La première solution envisageable pour la Dares est d'entrer en contact avec des sites diffuseurs d'offres et de créer des partenariats pour obtenir un accès privilégié à leurs données. Les partenariats ont l'avantage de garantir un accès régulier aux données pendant la durée du partenariat. Les données sont traitées et structurées en amont par les sites, avant d'être transmises à la Dares, qui peut facilement les exploiter et éventuellement solliciter des interlocuteurs pour obtenir une aide à leur compréhension. Toutefois, l'établissement de

[10] Les expertises *a priori* se basent sur l'affichage des sites. En pratique, il peut y avoir des divergences entre ce que les sites déclarent et ce qu'ils diffusent. Un site se déclarant « généraliste » peut ne pas l'être, par exemple car sa politique tarifaire attire en réalité plutôt certains recruteurs, bien que tous les types d'offres soient acceptés.

[11] Les sites ne sont pas nommés dans ce document.

[12] *Application Programming Interface* (Interface de Programmation Applicative) : ce sont des interfaces informatiques permettant de transmettre des données structurées.

partenariats pour de tels volumes de données et sans véritable contrepartie possible semble difficile à mettre en œuvre compte tenu du coût technique qu'il engendre pour les sites.

Les APIs, développées par certains diffuseurs comme Pôle emploi, LinkedIn, Indeed ou Glassdoor[13], permettent également de collecter facilement des données d'offres d'emploi. Ces APIs s'adressent principalement à des développeurs d'applications de recrutement et ont pour objectif (sauf dans le cas de Pôle emploi) de générer du trafic vers les sites diffuseurs. Elles sont faciles d'utilisation (mises à jour régulières, documentation…) et permettent d'accéder à des données structurées de qualité. Malheureusement, ces services ne sont pas à ce jour proposés par tous les sites et ne sont généralement pas accessibles gratuitement dans un but de collecte massive de données – les sites diffuseurs n'ayant *a priori* pas intérêt à partager l'intégralité de leurs contenus sans contrepartie.

La dernière possibilité identifiée est le *webscraping*, aussi appelé *scraping.* Le *scraping* consiste à collecter automatiquement des données depuis un site web. Cette solution nécessite un investissement technique initial : le développement et la mise en place de scripts de *scraping* automatiques et spécifiques à chaque site. Ensuite, une maintenance régulière est nécessaire car il faut adapter les *scrapers* à chaque changement de structure des sites. Ainsi, plus le nombre de sites augmente, plus les coûts de maintenance sont élevés. Enfin, il existe un fort risque de blocage de la part des sites, qui se protègent de plus en plus vis-à-vis de ces techniques d'extraction de données. Mais la mise en place initiale peut être relativement rapide et l'accès aux données a l'avantage de se faire sans intermédiaire : en récupérant quotidiennement ce qui a été publié par le site, le *scraping* donne une idée en temps réel des offres auxquelles n'importe quel utilisateur a accès. Pour ces raisons, et car les coûts en maintenance sont peu élevés pour le faible nombre de sites étudiés, la Dares a retenu cette solution.

c. Le *scraping* en pratique

Outils techniques et mise en œuvre à la Dares

La Dares a développé un outil *ad hoc* de *scraping* d'offres d'emploi en Python[14]. Des scripts de lancement automatique des *scrapers* sur un serveur distant dédié – une VM *(Virtual Machine)* – et de rapatriement des données sur des serveurs internes ont également été mis en place. Enfin, un outil d'alerte par mail en cas de défaillance des scripts a été élaboré. Ce code est régulièrement mis à jour afin de s'adapter aux changements de structure des sites.

Les *scrapers* sont spécifiques à chaque site et se composent de deux parties : les *crawlers* et les *parsers*. Chaque jour, les *crawlers* parcourent le site à la recherche des offres d'emploi et établissent la liste des urls des offres présentes sur le site. Les offres qui n'étaient pas en ligne la

[13] Les offres collectées par Pôle emploi sont accessibles par l'API Offres d'emploi v2 ; Indeed.com permet aux propriétaires de site web de rediffuser des offres d'emploi sur leur site grâce à l'API Job Search ; Glassdoor.com développe également des API accessibles à ses partenaires.

[14] Langage de programmation, Python 3.6, https://www.python.org/

veille (ou de manière générale, au précédent lancement du *scraping*) sont captées[15] (par requête HTTP). Ensuite, pour chaque offre d'emploi, les *parsers* sont chargés d'extraire les informations d'intérêt (métier, entreprise, lieu…) et de les enregistrer dans des fichiers csv. Le code html des offres est sauvegardé. Les offres qui ne sont plus en ligne par rapport à la veille sont archivées et leur date de disparition est enregistrée.

Aspect juridique

Le cadre juridique du *scraping* à visée statistique n'est pas bien défini à ce jour, bien que son existence soit reconnue par le CNIS[16], et il n'existe pas encore de politique commune vis-à-vis de cette pratique au sein des services statistiques publics français. La politique de référence est donc pour l'instant celle d'Eurostat, qui a défini en 2017 un code de bonne conduite[17], aussi appelé « nétiquette ».

Ce guide pratique est suivi à la Dares. Tout d'abord, la Dares prévient les sites par mail qu'ils vont faire l'objet d'une procédure de *scraping* quotidienne à des fins statistiques : en cas de refus formel d'un site, les travaux ne sont pas poursuivis sur ce site. Ensuite, la Dares veille à collecter les données lorsque les sites sont le moins fréquentés - la nuit - et les requêtes sont espacées d'un délai d'une seconde, afin de ne pas surcharger les serveurs. Les procédures de *scraping* mises en place minimisent le nombre de requêtes. Enfin, toutes les requêtes sont clairement identifiables comme provenant de la Dares et une adresse mail de contact est renseignée dans le 'user-agent'. Il est d'ailleurs dans l'intérêt de la Dares de respecter cette nétiquette car il est relativement simple pour les sites concernés d'interdire ou de limiter l'accès à leurs données. Finalement, la Dares a rencontré peu de problèmes avec les sites Internet scrapés. Convaincus de la finalité statistique du *scraping*, ils se sont montrés dans l'ensemble coopératifs et pour certains d'entre eux, intéressés par les recherches menées.

Performances d'extraction des offres

Pendant la période de collecte[18], l'activité des *scrapers* a pu être perturbée car, soit ils ont dû être arrêtés (mise à jour du code, maintenance de la VM), soit les codes n'ont pas fonctionné correctement et ils se sont arrêtés d'eux-mêmes.

Même lorsque les scrapers ont fonctionné correctement, certaines requêtes d'offres d'emploi n'ont pas abouti. Les principales causes d'échecs d'une requête sont l'interruption de service des serveurs cibles (sites en maintenance, code HTTP 503), l'interdiction d'accès de la part des sites (code HTTP 403) et la disparition de l'offre entre le moment où l'url est *crawlée* (i. e. repérée

[15] Certaines offres retirées, puis remises en ligne, ont ainsi pu être captées plusieurs fois : la liste des offres à récupérer est obtenue en comparant la liste des URL visibles au jour J par rapport aux URL visibles à J-1, et non à l'ensemble du stock d'URL récupérées depuis le lancement initial du *scraping.* Si une offre est publiée à J-2, retirée à J-1 et remise en ligne à J, elle est captée deux fois (à J-2 et J).

[16] Le *scraping* a notamment été discuté lors de la rencontre du CNIS, Les enjeux des nouvelles sources de données, juillet 2018

[17] Eurostat, ESSnet Big Data WP2 Deliverable 2.1, Annex 1, février 2017

[18] La période de collecte considérée va du 1er janvier au 31 décembre 2019 pour les *scrapers* lancés fin 2018 ou du jour de lancement au 31 décembre 2019 pour les *scrapers* lancés courant 2019.

comme étant en ligne) et l'émission de la requête HTTP (code HTTP 410 ou 404). Ainsi, en 2019, 98,6 % des requêtes d'url d'offres d'emploi ont réussi (code HTTP 200).

Ces informations sur la performance des outils développés ne reflètent pas exactement le taux d'offres captées par rapport au nombre d'offres réellement publiées sur le site. D'un côté, une offre publiée pendant une période d'arrêt des *scrapers* sera captée au moment de la reprise si elle est toujours en ligne (ce qui est le cas d'un certain nombre d'offres, étant donné la longévité des offres en ligne). D'un autre côté, le *scraping* étant journalier, une offre qui apparaît et disparaît le jour-même n'est pas captée.

Finalement, bien que des améliorations soient possibles, le *scraping* est considéré comme satisfaisant. Cet outil est d'ailleurs amené à évoluer et fait l'objet d'une maintenance continue.

2. D'une donnée brute à une donnée statistique

a. Harmonisation des données

Bien que les offres d'emploi répondent à une structure commune assez normée, la quantité d'informations qu'elles contiennent est variable selon les sites : certains proposent des offres très détaillées, alors que d'autres imposent un format plus concis. Ainsi, une offre doit contenir *a minima* les informations structurées[19] suivantes pour être retenue : intitulé du métier, lieu de travail (code postal ou, à défaut, départemental)[20] et nom de l'employeur. Ces variables ont été choisies car elles sont disponibles sur la plupart des offres, correspondent aux champs nécessaires à l'étape de déduplication (cf. partie II.2.c) et sont utiles dans le cadre des analyses. En 2019, 9 % des offres scrapées sont retirées du champ car elles ne contiennent pas toutes les variables requises[21]. Le taux d'offres inutilisables pour cette raison varie de 0,1 % à 12 % selon les sites, à l'exception d'un site dont le taux de rejet atteint 23 % en 2019[22].

Sur certains sites, les données structurées recueillies sont beaucoup plus riches que ces quatre champs incontournables : secteur d'activité, temps de travail, salaire, SIREN ou SIRET… Toutefois, des différences de nomenclature ou de granularité entre les sites rendent difficile l'exploitation de ces variables.

b. Recoder le métier : une approche par *machine learning*

Si les variables lieu de travail et nom de l'employeur sont utilisées telles quelles, un travail plus approfondi a été mené sur l'intitulé du métier. En effet, l'intégration des offres en ligne aux travaux de la Dares nécessite d'identifier le métier de l'offre. Plus, précisément, de classer correctement les offres dans une nomenclature officielle de métiers, le répertoire opérationnel des métiers et des emplois (ROME).

---

[19] Est désignée par information structurée une donnée contenue dans un champ clairement délimité de la page html et faisant référence à une seule information : par exemple, un code postal contenu dans une balise html propre est une information structurée mais le descriptif (texte libre) de l'offre n'en est pas une.

[20] Certaines offres ne comportant qu'un nom de ville ont pu être recodées par appariement avec une base de codes postaux. Les offres pour des emplois à l'étranger sont écartées.

[21] *A contrario*, des offres qui ne proposent pas un contrat de travail salarié (missions *free-lance*, adhésion à une plateforme…) peuvent être retenues.

[22] Ce taux de rejet élevé est dû à un problème de récupération du lieu de travail et du nom de l'entreprise dans certaines offres d'emploi. Ce problème est résolu pour les données collectées les années suivantes.

Un algorithme de *machine learning* a été utilisé pour classer les offres par métier. Cet algorithme – appelé *Support Vector Machine* (SVM) – est un algorithme d'apprentissage supervisé : il « apprend » à classer les offres à partir d'exemples d'offres d'emploi dont le métier est connu. Après avoir été entraîné sur des offres annotées en code ROME, il est capable de prédire (avec un certain taux d'erreur) le métier d'une offre. Les étapes nécessaires à l'entraînement de cet algorithme de *machine learning* sont détaillées ci-dessous : collecte des données annotées qui serviront d'exemples, nettoyage des données, entraînement de l'algorithme et évaluation des performances.

Présentation du problème de classification

Le Répertoire opérationnel des métiers et des emplois (ROME) est la nomenclature des métiers établie par Pôle emploi. Il contient 532 métiers répartis en 110 domaines professionnels, eux-mêmes regroupés en 14 grands domaines (Tableau 3). Cette nomenclature recense pour chaque code ROME des exemples d'appellations métier (11 081 au total).

TABLEAU 3 - Nomenclature ROME

Arborescence	Nombre de classes	Exemple
Grands domaines	14	H (Industrie)
Domaines professionnels	110	H12 (Conception, recherche, études et développement)
Code ROME	532	H1204 (Design industriel)
Appellation	11 081	Designer / Designeuse automobile

Source : Pôle emploi, Répertoire Opérationnel des Métiers (ROME).

Le codage du métier est donc un problème classique d'analyse textuelle : il faut classer des chaînes de caractères (les intitulés de métier recueillis dans les offres en ligne) dans les classes de la nomenclature (les codes ROME). Par exemple, l'intitulé « Designer automobile débutant – CDI – Paris » devra être classé en H1204 (Design industriel).

Présentation des données utilisées pour l'apprentissage

Afin de pouvoir entraîner les modèles et évaluer leurs performances, environ 1 547 800 offres collectées par Pôle emploi annotées en code ROME ont été recueillies entre le 23 janvier et le 31 décembre 2019 *via* une API de Pôle emploi[23]. Pour chacune de ces offres, le titre (ou intitulé métier) et le code ROME associé sont extraits. À ces 1 547 800 couples intitulés-code ROME sont ajoutées les appellations ROME renseignées dans l'arborescence officielle du ROME. Finalement, 1 558 800 intitulés de métier avec leur code ROME correspondant ont été obtenus.

[23] https://www.emploi-store-dev.fr/portail-developpeur-cms/home/catalogue-des-api/documentation-des-api/api/api-offres-demploi-v2.html

Nettoyage des libellés de métier

Afin de faciliter l'apprentissage, les données sont nettoyées. En effet, les titres d'offres d'emploi contiennent parfois d'autres informations que le métier (type de contrat, lieu …) qui peuvent perturber les modèles. Des traitements habituels d'analyse textuelle[24] leur sont donc appliqués :

- Suppression des caractères spéciaux, de la ponctuation et des chiffres, conversion en minuscules, suppression des mots vides[25] (tels que « le », « la », « de », …),
- Suppression des mots relatifs au type de contrat, à la qualification ou à la quotité de temps de travail (par exemple « CDI » ou « temps plein »),
- Suppression des noms de pays, régions et départements français,
- Harmonisation des noms de métiers au masculin à l'aide d'une liste de métiers déclinés au masculin et au féminin établie à partir du ROME et des données collectées (par exemple, « boulangère » devient « boulanger »),
- Explicitation des acronymes et abréviations (« DRH » devient « directeur des ressources humaines », « prof » devient « professeur »),
- Lemmatisation des mots à l'aide du lexique Morphalou (ATILF, 2019), afin de réduire la taille du vocabulaire,
- Suppression des mots en double.

Modélisation des libellés de métier

Une fois nettoyés, les libellés de métier sont transformés en vecteurs : c'est un préalable à l'entraînement d'un modèle de type *machine learning*. Pour cette étude, les libellés de métiers sont modélisés par des vecteurs de type « sac de mots » binaires. Autrement dit, les libellés sont représentés par des vecteurs de la taille du vocabulaire (~22 000 mots) et, pour chaque libellé, chacun des mots du vocabulaire prend la valeur 1 ou 0 selon si le mot est présent dans le libellé ou non[26].

Choix d'un algorithme de machine learning

Après avoir testé une méthode de classification par calcul de distance textuelle sur un jeu de données restreint (DE MARICOURT, 2018), la collecte d'un grand volume de données annotées a permis d'envisager l'utilisation d'un algorithme d'apprentissage supervisé. En 2018, BOSELLI et *al.*, ont réalisé une comparaison d'algorithmes de *machine learning* pour un problème similaire[27]. Sans avoir pu reproduire l'ensemble de ces travaux sur nos données, nous avons retenu la classe d'algorithmes la plus performante selon les auteurs : les machines à vecteurs supports (SVM). Compte tenu de nos capacités de calcul et de la grande taille de notre jeu de données, nous avons choisi à ce stade la configuration la moins intensive en calcul, un SVM linéaire (ayant comme paramètre de régularisation C=1). De manière classique en *machine learning*, nous avons divisé aléatoirement notre jeu de donnés en un jeu d'entraînement (80 % du jeu de données) et un jeu de test (20 % du jeu de données), puis entraîné notre modèle sur le premier jeu. Sur ces données et en utilisant la librairie Python Scikit-learn (PEDREGOSA et *al.,* 2011*),* le modèle est entraîné en

[24] L'ensemble du code (nettoyage, lemmatisation, classification) et des ressources (listes de mots) sont disponibles sur https://github.com/OnlineJobVacanciesESSnetBigData/JobTitleProcessing_FR.

[25] Les mots vides sont des mots très fréquents qui n'apportent pas beaucoup de sens. Ils sont aussi appelés *stopwords*.

[26] L'ordre des mots dans un libellé n'est donc pas pris en compte.

[27] BOSELLI et *al.* ont étudié la classification multilingue d'offres d'emploi à partir de leurs descriptifs.

moins d'une heure. Bien que performante, cette méthode de classification pourrait être améliorée, en utilisant le texte de l'offre (en plus du titre), en optimisant les hyperparamètres du modèle ou en ayant recours à d'autres modèles d'apprentissage supervisé.

Résultats

Le modèle obtenu après entraînement est testé sur les 20 % restant du jeu de données[28]. Différentes métriques sont utilisées pour évaluer les performances du modèle :

- l'*accuracy*, qui est le taux de libellés dont le code ROME (respectivement domaines professionnels ou grands domaine) prédit correspond au code ROME réel ;
- le score F1, qui mesure la performance de l'algorithme sur chaque code ROME (respectivement domaines professionnels ou grands domaines) et dont nous calculons la moyenne pondérée[29] par code ROME (respectivement domaines professionnels ou grands domaines) et la moyenne non pondérée.

Ces métriques sont calculées aux trois niveaux de la nomenclature ROME (Tableau 4).

TABLEAU 4 - Performances de la classification des offres par code ROME sur le jeu de test

Niveau de la nomenclature	*Accuracy*	F1-score *Moyenne pondérée*	F1-score *Moyenne non pondérée*
Grands domaines	0,97	0,97	0,96
Domaines professionnels	0,96	0,96	0,93
Code ROME	0,94	0,94	0,91

Lecture : le modèle de classification prédit correctement le grand domaine de 97 % des données de test. En moyenne par code ROME, le score F1 de la prédiction est de 0,91.
Champ : ensemble du jeu de test.
Source : Dares.

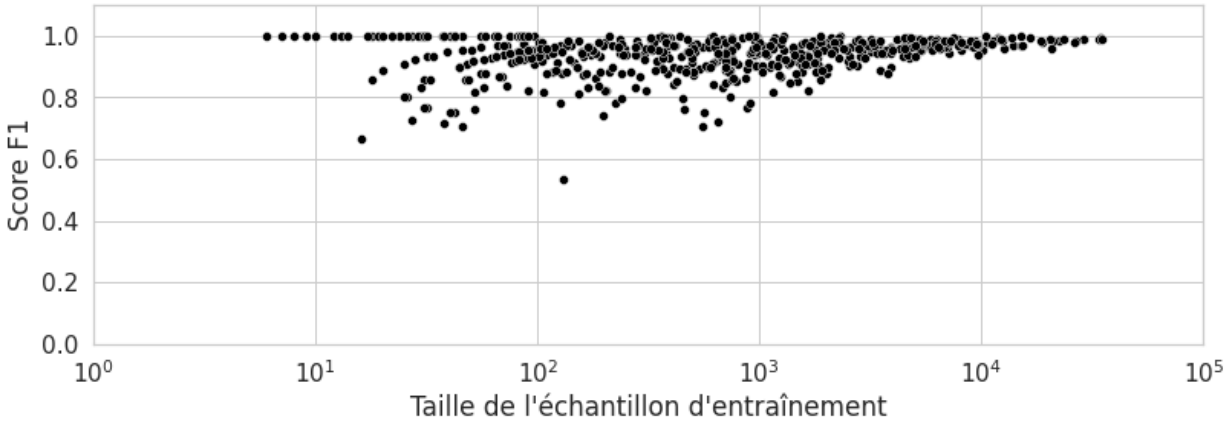
Si le modèle prédit assez bien les grands domaines (*accuracy* de 0,97 ; F1-score non pondéré de 0,96), il est moins performant à des niveaux plus détaillés : au niveau du code ROME, l'*accuracy* n'est plus que de 0,94 et le F1-score non pondéré de 0,91. L'écart entre le score F1 non pondéré et le score F1 pondéré au niveau fin est lié au fait que le modèle prédit bien les codes ROME les plus représentés dans le jeu de données au détriment des plus rares[30] (Figure 3). Toutefois, avec un score F1 non pondéré de 0,91, le *classifier* n'est pas « naïf » (i. e. il ne se contente pas de prédire uniquement les classes les plus représentées).

[28] Les performances sont mesurées sur le jeu de test plutôt que le jeu d'entraînement afin d'évaluer le modèle sur des données qu'il ne « connaît » *a priori* pas (bien que certains intitulés d'offres redondants dans le jeu de données se retrouvent à la fois dans le jeu d'entraînement et dans le jeu de test).

[29] La pondération est effectuée selon la distribution du jeu de test créé à partir des offres de Pôle emploi.

[30] Les cinq codes ROME les moins bien prédits sont : les métiers de la direction d'exploitation en assurances (F1 score de 0,53), de la manutention portuaire (0,67), de la mise en oeuvre et du pilotage de la politique des pouvoirs publics (0,80), du management en exploitation bancaire (0,71) et du contrôle et de l'inspection du Trésor Public (0,71).

FIGURE 3 - Distribution des scores F1 par code ROME selon la taille de l'échantillon d'entraînement



Lecture : les codes ROME ayant plus de 1 000 observations dans leur échantillon d'entraînement ont un score F1 supérieur à 0,8. Un seul code ROME avec plus de 100 données d'entraînement a un score F1 inférieur à 0,6.
Champ : ensemble du jeu de test.
Source : Dares.

La dissimilarité de Duncan entre les jeux de test prédit et réel au niveau des codes ROME est faible, à 1,4 %. Mais les variations d'effectifs par métier liées à l'utilisation de ce modèle ne sont pas aléatoires. Les effectifs prédits sont surestimés par rapport aux effectifs observés (+36,8) pour les classes très fréquentes - avec plus de 10 000 exemples d'entraînement - au détriment des classes plus rares (Tableau 5). La structure des codes ROME prédits est ainsi moins dispersée que la répartition observée des ROME, avec une surreprésentation des quelques classes avec un large nombre d'exemples : 66 % des codes ROME avec plus de 10 000 exemples sont surreprésentés dans la prédiction contre 7 % de ceux ayant 100 exemples ou moins. Malgré leurs très bons scores d'un point de vue *machine learning*, les codes ROME avec beaucoup d'exemples contribuent plus à la dissimilarité.

TABLEAU 5 - Différences entre effectifs prédits et observés par ROME regroupés par taille d'échantillons d'entraînement

Taille de l'échantillon d'entraînement	Nombre de codes ROME	Différence moyenne (en effectifs)	Surreprésentation ou sous-représentation des codes ROME dans la prédiction (%)		
			Surreprésentation	Egalité	Sous-représentation
Inférieur ou égal à 100	109	-1,0	7	41	51
Entre 100 et 1 000	213	-5,6	23	6	71
Entre 1 000 et 10 000	181	1,3	39	2	59
> 10 000	29	36,8	66	3	31

Lecture : en moyenne et pour les codes ROME dont la taille de l'échantillon d'entraînement est comprise entre 100 et 1 000, les effectifs prédits du jeu de test sont inférieurs de 5,6 aux effectifs « réels ». 71 % des codes ROME dont la taille de l'échantillon d'entraînement est entre 100 et 1 000 sont sous-représentés dans la prédiction.
Champ : ensemble du jeu de test.
Source : Dares.

Ce modèle présente d'autres limites pour son application aux offres collectées en ligne. D'une part, la formulation des métiers peut différer entre les offres d'emploi en ligne et celles déposées à Pôle emploi. D'autre part, la structure peut fortement varier entre les données d'entraînement et

les offres collectées. Ces deux éléments pourraient rendre le modèle moins performant sur les données scrapées. Pour évaluer ces possibles effets, il faudrait annoter manuellement un échantillon d'offres collectées en ligne. Ce test, très chronophage, n'a pour l'instant pas été tenté.

c. Déduplication des offres d'emploi

Une fois dans un format commun, les offres d'emploi sont dédupliquées. Deux méthodes sont utilisées pour identifier les doublons : la déduplication par appariement des URL et la déduplication par proximité textuelle. Les doublons peuvent être des offres publiées plusieurs fois sur un même site (ce sont les doublons « intrasites ») ou des offres publiées sur plusieurs sites différents (doublons « intersites »).

La déduplication par appariement des URL consiste à identifier les URL présentes plusieurs fois dans la base et à n'en compter qu'une occurrence. Elle permet d'identifier des doublons intrasites et repose sur l'hypothèse qu'une URL correspond à une seule offre.

La deuxième déduplication, par similarité textuelle, s'opère en deux étapes. La première étape consiste à regrouper les offres ayant le même code ROME, le même nom d'entreprise et le même lieu de travail (au niveau du code postal ou du code départemental, selon l'information disponible) qui ont été publiées à 14 jours de différence au maximum. À l'intérieur de chaque groupe, les offres sont comparées deux à deux : si leurs descriptifs (ou corps de l'offre) ont des vocabulaires semblables à plus de 95 % (ou plus exactement des vocabulaires ayant une similarité au sens de Jaccard[31] supérieure à 0,95), elles sont considérées comme des doublons ; alternativement, elles sont supposées distinctes.

La déduplication par similarité textuelle dépend donc de deux paramètres : la fenêtre temporelle de déduplication (14 jours) et le seuil de proximité textuel (0,95). À ce stade, ces paramètres ont été fixés de manière arbitraire. Si ces paramètres permettent de détecter 14 % de doublons intrasites (Figure 4), seul 1 % des offres scrapées sont identifiées comme des doublons intersites. Bien qu'il soit difficile d'estimer la part réelle de doublons intersites, ce taux semble assez faible[32]. Les paramètres pourront donc être amenés à évoluer, voire être modulés en fonction de la provenance des offres. Par exemple, le seuil de proximité textuelle pourrait être abaissé si les offres proviennent de sites différents ou rehaussé pour des offres issues d'une même source (pour des raisons de design du site ou de public visé, la formulation d'une même offre peut varier d'un site à un autre).

Au-delà du choix des paramètres, ce procédé peut entraîner une surestimation des doublons. En effet, les offres qui sont des doublons sur la forme mais qui correspondent en fait à plusieurs postes à pourvoir sont éliminées. Si un employeur diffuse plusieurs offres identiques (même métier, même lieu de travail, même description) pour plusieurs postes différents, une seule offre sera conservée.

---

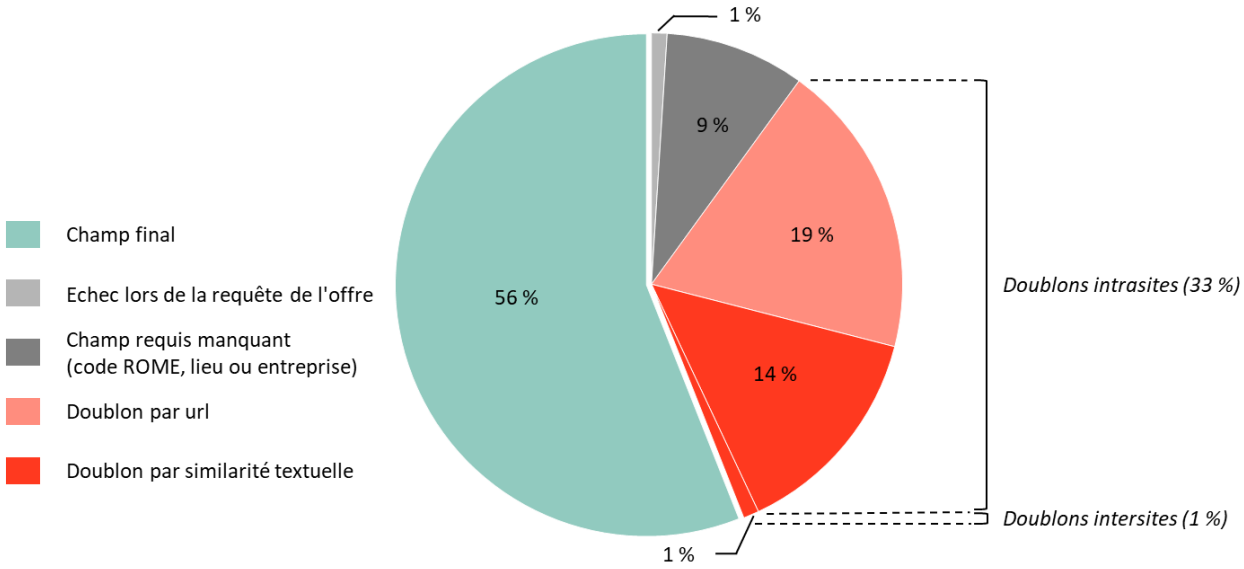
[31] La similarité - ou indice - de Jaccard permet de calculer la proximité entre deux ensembles. Ici, les ensembles considérés sont les vocabulaires (V_1et V_2) utilisés dans les descriptifs des deux offres à comparer. L'indice de Jaccard correspond au cardinal de l'intersection des vocabulaires divisé par le cardinal de leur union, soit : $\left|\frac{V_1 \cap V_2}{V_1 \cup V_2}\right|$

[32] À titre de comparaison, Pôle emploi trouvait un taux de doublons de moins de 4 % dans son agrégateur en décembre 2015 : https://www.pole-emploi.org/accueil/actualites/infographies/lagregation-des-offres-demploi.html?type=article#.

Finalement, 34 % des offres scrapées en 2019 sont identifiées comme étant des doublons (Figure 4), dont une grande majorité de doublons intrasites (33 %). Les doublons intrasites par URL (19 % des offres scrapées) sont en partie générés par des dysfonctionnements des outils de *scraping*[33] et il est difficile de distinguer les republications réelles des bugs des *scrapers*. Concernant la similarité textuelle, l'analyse de quelques doublons intrasites donne deux pistes d'explication. Les doublons analysés correspondent soit à des republications après une longue période en ligne, soit à des successions de publications très courtes. Les republications après une longue période en ligne – généralement 30 jours – peuvent être liées aux délais d'expiration des offres sur les sites d'emploi en ligne : quand l'offre est fermée automatiquement par le site, l'employeur n'ayant pas trouvé de candidat adéquat la publie à nouveau. Les publications successives d'offres sur de petites périodes semblent se produire pour des offres qui n'ont pas été déposées directement sur le site. En effet, de nombreux sites publient des offres provenant de sites tiers (cf. partie I.3). La succession des publications pourrait alors s'expliquer par des redirections de trafic d'un site à un autre, sans déduplication par le site agrégateur. Ces derniers éléments suggèrent que le phénomène de republication d'une offre dépend de l'offre et de l'employeur mais également du site.

À l'issue de la sélection et de la déduplication des données, la base d'exploitation statistique Jocas représente 56 % des offres scrapées en 2019 (Figure 4), soit un peu plus de 5 millions d'offres.

FIGURE 4 - Répartition des offres scrapées en 2019



Lecture : parmi les offres scrapées, 10 % ne sont pas exploitables et 34 % sont des doublons. La base Jocas 2019 contient 56 % de l'ensemble des offres scrapées en 2019.
Champ : ensemble des offres d'emploi scrapées par la Dares en 2019.
Source : Dares.

[33] Par exemple, si la mise à jour de l'historique des offres en ligne et scrapées ne se fait pas correctement, une offre déjà scrapée peut l'être à nouveau. Cela se produit notamment si le scraper rencontre un problème et s'interrompt avant l'étape de mise à jour des URL scrapées.

d. Passage de la nomenclature des métiers à celles des familles professionnelles

Afin de comparer les données aux autres sources de la Dares, il est nécessaire de convertir les offres dans la nomenclature des familles professionnelles (FAP). Cette opération s'effectue traditionnellement en appliquant une matrice de passage nécessitant de disposer de la qualification du poste. Or, cette dernière n'est pas disponible sur tous les sites. C'est la raison pour laquelle une autre table de passage permettant de basculer directement des ROME aux FAP a été mobilisée. Cette table est obtenue à partir des croisements ROME x FAP observés sur le champ des offres déposées à Pôle emploi en 2019 et des offres scrapées contenant la variable de qualification.

III. Les caractéristiques de la base d'offres en ligne de la Dares (Jocas 2019)

Cette section présente dans un premier temps les principales caractéristiques de la base Jocas 2019 de la Dares. Dans un second temps, elle fournit un cadrage de cette source statistique par rapport aux sources classiques du marché du travail, en termes de couverture des métiers et du territoire. Enfin, la dernière partie présente des cas d'usage de la base Jocas.

1. Composition de la base d'offres en ligne Jocas 2019

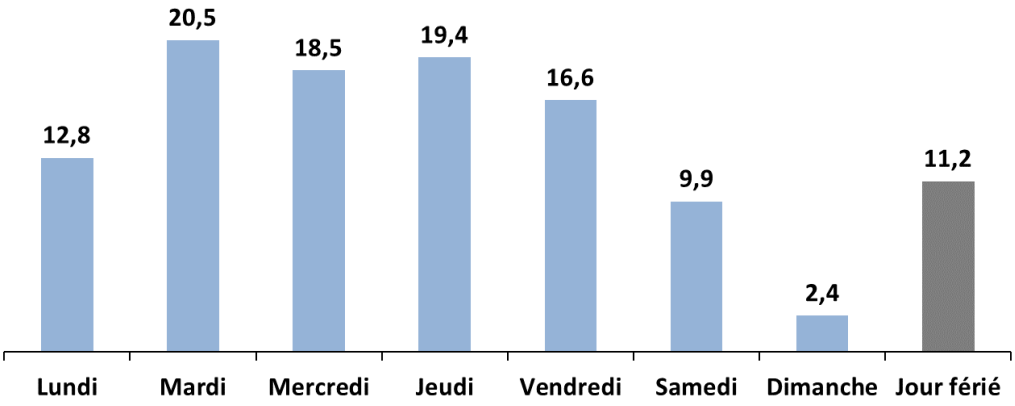
a. Temporalité de la source Jocas

Concernant l'aspect temporel, une date de publication a pu être identifiée[34] pour 96 % des offres de la base Jocas 2019. Cependant, pour l'année 2019, en raison de la montée en charge de l'outil, le nombre d'offres d'emploi publiées en ligne par mois ou par semaine ne semble pas refléter une réalité saisonnière. Par exemple le nombre d'offres augmente en fin d'année car deux sites ont été ajoutés au quatrième trimestre 2019. Au contraire, un « bug » de l'outil de *scraping* a fait chuter le nombre d'offres correctement collectées en juin 2019.

Par ailleurs, au niveau infra-hebdomadaire, le nombre moyen d'offres publiées par jour varie fortement selon le jour de la semaine, avec un pic de publication le mardi (20,5 % des offres d'une semaine sont publiées le mardi, Figure 5) et le jeudi (19,4 %). Sans surprise, le dimanche est le jour qui compte en moyenne le moins d'offres publiées en 2019 (2,4 %), loin derrière le samedi (9,9 %) ou un jour férié (11,2 %). Ce plus grand nombre d'offres les samedis et jours fériés par rapport au dimanche pourrait en partie s'expliquer par les protocoles de validation des annonces parfois mis en place par les sites. En effet, certains sites vérifient la conformité du contenu des offres avant de les mettre en ligne. Cela peut entraîner un délai entre le moment où le recruteur crée l'annonce et la date effective de publication : par exemple une offre renseignée le vendredi par un employeur peut être mise en ligne seulement le samedi, après validation par le site. Ce décalage d'un jour sur l'autre peut aussi expliquer le plus faible nombre d'offre le lundi par rapport aux autres jours ouvrés de la semaine.

[34] La date de publication est distinguée de la date de scraping : la date de scraping est la date à laquelle l'offre a été scrapée, alors que la date de publication est la date à laquelle elle a été publiée sur le site. La date de publication est soit identique, soit antérieure à la date de scraping.

FIGURE 5 – Répartition des offres Jocas 2019 sur une semaine



Note : la répartition des offres sur une semaine est calculée sur les semaines complètes, qui n'ont pas de jour férié. La part d'offres publiées sur un jour férié est calculée sur les semaines complètes avec un jour férié.

Lecture : en moyenne en 2019, sur une semaine sans jour férié, 20,5 % des offres de la semaine sont publiées le mardi.

Champ : ensemble des offres Jocas pour lesquelles une date de publication en 2019 a été identifiée.

Source : Jocas 2019, Dares.

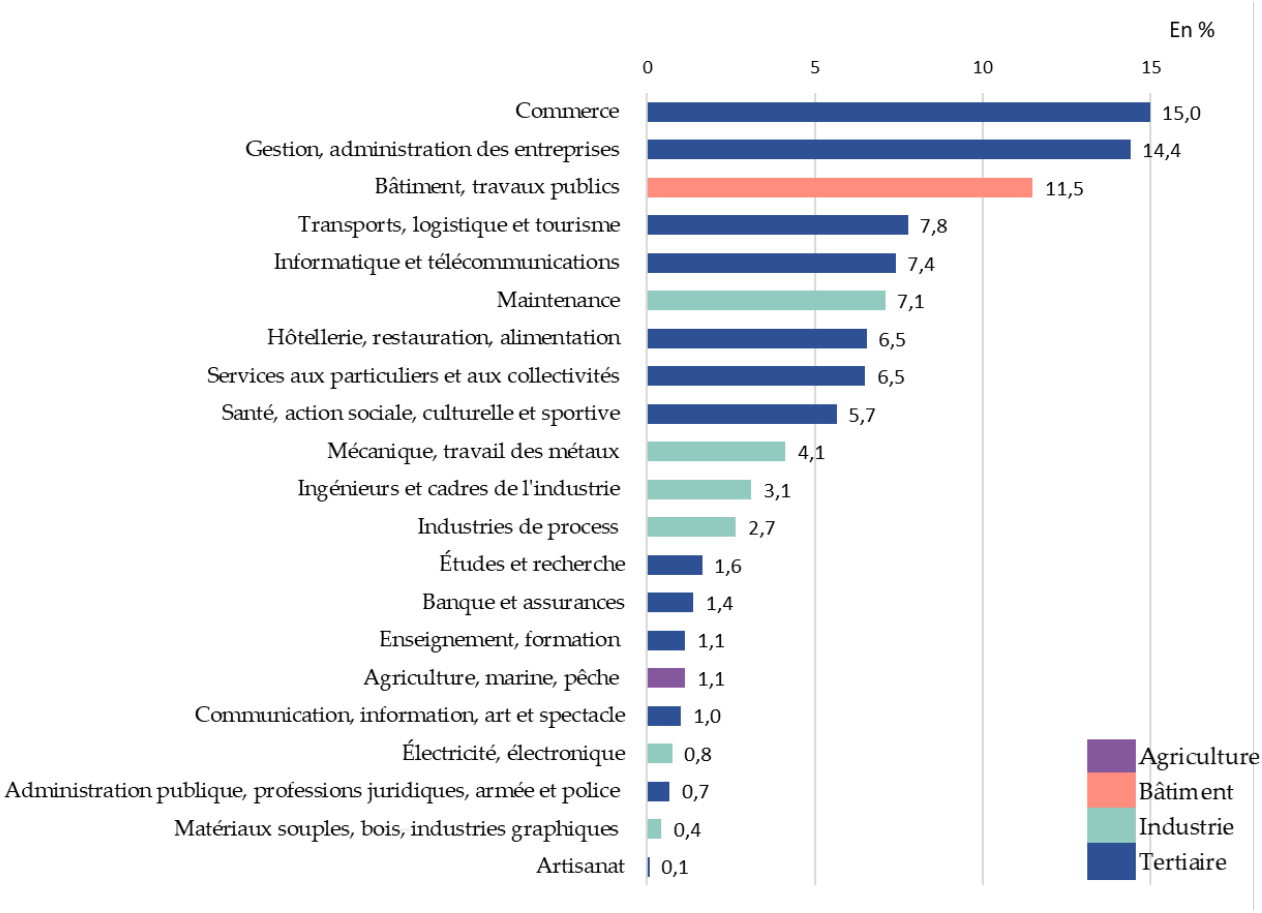
b. Composition des offres par domaine professionnel et par région

La base Jocas 2019 est constituée de plus de 5 millions d'offres dédupliquées et exploitables (à comparer aux 3,3 millions directement collectées par Pôle emploi en 2019[35]). Pour ces offres, le métier et le lieu de travail (au niveau départemental) sont connus : cette première partie détaille la composition de la base Jocas selon ces informations.

Au niveau des métiers, 21 des 22 domaines professionnels sont couverts, à l'exception de celui, attendu, de la politique et de la religion. Les métiers de service représentent sept offres en ligne sur dix dans la base Jocas 2019, notamment dans les domaines professionnels du commerce (15 %) et de la gestion et administration des entreprises (14 %). Par ailleurs, respectivement 18 % et 11 % des offres Jocas sont diffusées dans les grands domaines de l'industrie et du bâtiment (Figure 6).

[35] Les informations sur le volume d'offres d'emploi collectées (ainsi que diffusées) par Pôle emploi peuvent être retrouvées sur https://statistiques.pole-emploi.org/offres/offres?fk=C&ss=1

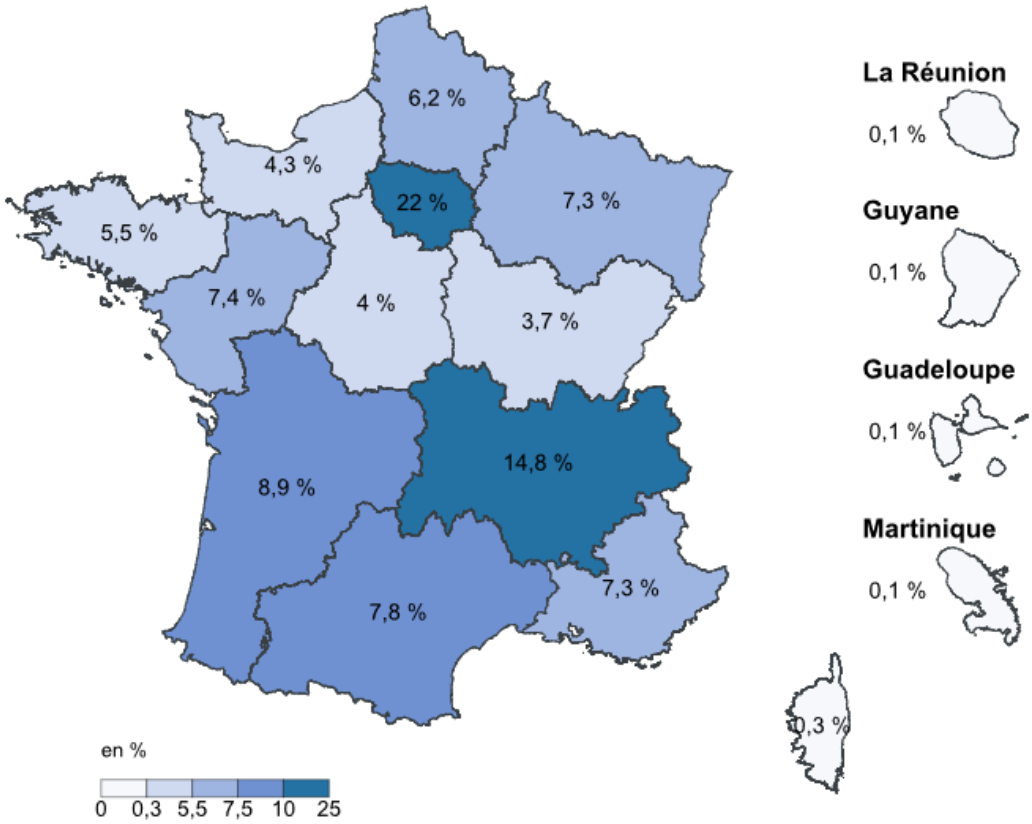
FIGURE 6 - Répartition des offres Jocas 2019 par domaine professionnel



Lecture : 15% de l'ensemble des offres Jocas proposent un poste dans le domaine professionnel du commerce.
Champ : offres scrapées auprès d'une quinzaine de sites diffuseurs d'offres d'emploi en ligne.
Source : Jocas 2019, Dares.

En ce qui concerne le lieu de travail, les offres Jocas couvrent l'ensemble du territoire national. Plus de deux offres sur dix (22 %, Figure 7) sont localisées en Île-de-France. La région Auvergne Rhône-Alpes, deuxième région pourvoyeuse, totalise 15 % du volume national d'offres. La part des autres régions est comprise entre 4 % et 9 %, à l'exception de la Corse et des régions d'outre-mer dont les parts sont inférieures à 1 %.

FIGURE 7 - Répartition des offres Jocas 2019 par région



Lecture : 22 % des offres Jocas sont pour des postes localisés dans la région Île-de-France.
Champ : offres scrapées auprès d'une quinzaine de sites diffuseurs d'offres d'emploi en ligne.
Source : Jocas 2019, Dares.

2. Couverture et représentativité des données

Au-delà de ce premier portrait, il semble nécessaire de positionner plus précisément les offres en ligne Jocas par rapport à l'ensemble des offres (en ligne ou non) et par rapport aux sources usuelles du marché du travail.

a. Être représentatif ou ne pas être ?

De nombreuses études effectuées à partir d'offres en ligne n'évoquent pas ou peu les problèmes de représentativité de celles-ci. Sans analyse de représentativité, les résultats obtenus à partir de telles données risquent d'être implicitement considérés comme représentatifs de l'ensemble des offres en ligne voire généralisés à l'ensemble des offres sans preuve (FABERMAN et KUDLYAK

2016). La représentativité des offres en ligne de la base Jocas est donc une des premières questions qu'il convient d'aborder avant son usage statistique.

Or, il n'existe ni enquête ni *a fortiori* de recensement du nombre total des offres d'emploi, qui pourrait servir de référence. Il n'est donc pas possible de comparer voire de caler les offres Jocas, comme cela peut être fait dans des enquêtes volontaires auprès des salariés (par exemple l'enquête *Wage indicator*) sur le web (KUREKOVA et *al*., 2014 ; LENAERTS et *al*., 2016). Il n'est pas non plus possible de traiter les offres non couvertes comme des données manquantes et d'utiliser des outils statistiques traitant les données manquantes, comme les modèles de probabilité prédisant des quantités à partir de données avec des valeurs non observées (ROYALL, 1992 ; LITTLE et RUBIN, 2002 ; LONGFORD, 2006).

Faute de cette information, quelques études ont testé plusieurs approches pour traiter les problèmes liés à la (non-)représentativité des offres en ligne. L'approche la plus répandue consiste à se servir de données représentatives décrivant la structure du marché du travail telles que les enquêtes Emploi (STEPHANIK, 2012 ; JACKSON, 2007). Ces dernières ne sont toutefois pas considérées comme de bons proxies pour mesurer la structure de la demande de travail, car elles incluent les deux volets du marché (offre et demande) ainsi que leur appariement (KUREKOVA et *al*., 2014).

En France, à défaut de données de référence sur l'ensemble des offres, on sait, d'après l'enquête Ofer 2016 de la Dares, que 49 % seulement des recrutements ont fait l'objet d'une annonce sur Internet (BERGEAT et *al*., 2018) et qu'ils ne sont pas représentatifs de l'ensemble des recrutements : le recours au canal Internet varie fortement selon les métiers (cf. partie I.3). Internet est souvent utilisé comme un canal complémentaire aux canaux prioritaires des recruteurs tels que les relations professionnelles (BERGEAT et *al.*, 2018 ; AVENTUR et al., 2017). Il est donc plutôt utilisé quand les recruteurs veulent procéder à une recherche extensive de candidats, généralement dans le cas des postes difficiles à pourvoir (BERGEAT et REMY, 2017). Ainsi, les offres en ligne renseignent davantage sur les professions pour lesquelles les recruteurs éprouvent d'importantes difficultés de recrutement en passant par les canaux traditionnels, que sur les emplois vacants de manière générale (KUREKOVA et *al*., 2014).

Certaines études visent un segment du marché du travail où la question de la couverture peut paraître moins cruciale. C'est par exemple le cas d'études qui porteraient sur les métiers du numérique, pour lesquels les offres d'emploi sont davantage publiées sur internet. D'autres études se basent sur des enquêtes sur les emplois vacants représentatives de l'ensemble des établissements (VAN OURS et RIDDER, 1992). D'autres encore utilisent des données administratives ou des enquêtes statistiques en complément aux offres postées en ligne, afin d'améliorer leur représentativité (MARTIKAINEN, 2010). Une correction de biais de représentativité peut aussi être effectuée en essayant de contrôler les caractéristiques des entreprises qu'on peut relier aux offres en ligne ou en mesurant la part de marché et les avancées techniques des sites diffuseurs d'offres en ligne dans le pays pour voir s'il existe un site dominant (KUREKOVA et *al*., 2014).

Finalement, en l'absence de source – administrative ou d'enquête – de référence sur l'ensemble des offres ou sur les offres en ligne, l'usage statistique de Jocas ne peut reposer sur sa stricte représentativité, qu'il n'est pas possible de vérifier. Il n'en reste pas moins que Jocas est source de nouvelles informations et notamment d'indicateurs sur le fonctionnement du marché du travail en ligne.

b. Positionnement de Jocas par rapport aux autres sources

Si l'absence d'une source de référence, couvrant toutes les offres d'emploi sur le marché du travail français, ne permet pas de qualifier précisément la couverture et la représentativité des offres Jocas, il est cependant possible de confronter ces offres aux autres sources disponibles sur le marché du travail (cf. partie I.2). Cette partie présente donc une comparaison en structure (par métier et par lieu) des offres Jocas avec les données des offres diffusées par Pôle emploi (y compris offres partenaires), des projets de recrutements (issus de BMO), du nombre d'emplois ayant débuté depuis moins d'un an suivant l'enquête Emploi[36], des intentions d'embauche (DPAE), des emplois vacants d'Acemo et des embauches de MMO (Tableau 1). Ces trois dernières sources, qui n'intègrent pas de dimension « métier » ont été déclinées par domaine professionnel (FAP 22) à l'aide d'une table de passage secteurs-métiers établie à partir des nombres d'emplois ayant débuté depuis moins d'un an de l'enquête Emploi en continu sur la période 2015-2019[37]. Ces sources, de grande taille, permettent un suivi *a minima* annuel et fin par métier du marché du travail. En revanche, elles ne distinguent pas la médiation par une offre en ligne sur le marché du travail.

Couverture des métiers

L'indice de dissimilarité Duncan est calculé pour mesurer la dissimilarité entre la structure des métiers dans la nomenclature des familles professionnelles agrégée en 22 domaines professionnels (FAP 22) des offres Jocas et celle de chacune des autres sources (Tableau 6). Les offres Jocas sont plus proches des offres diffusées par Pôle emploi (*i. e.* les offres collectées par l'opérateur *stricto sensu* et issues de la statistique du marché du travail – STMT–, ainsi que les offres partenaires) selon cet indice qui atteint alors 10, sa valeur minimale. Autrement dit, la redistribution de seulement 10 % des offres Jocas permet de retrouver une structure semblable à celle des offres diffusées en ligne par Pôle emploi. Cela n'est pas surprenant : parmi l'ensemble des sources de référence retenues, les offres diffusées par Pôle emploi sont celles qui ressemblent le plus par leur nature aux offres Jocas. La seconde source la plus proche est celle des emplois vacants d'Acemo (16 %), suivie par les nombres d'emplois de moins d'un an issus de l'enquête Emploi en continu (EEC) de l'Insee (21 %) ; Jocas est nettement plus éloignée des trois autres sources que sont BMO, DPAE et MMO (indices de Duncan à 32 %, 30 % et 29 % respectivement).

La source Ofer aurait également pu servir de point de comparaison. Elle permet - contrairement aux autres sources - une comparaison sur les recrutements donnant lieu à une diffusion d'offre en ligne. Cependant, à cause de son champ de métiers plus restreint et d'une plage temporelle en décalage (enquête menée en 2015), elle n'est pas retenue ici (Annexe 2).

[36] Mesurés par les emplois avec une ancienneté inférieure à un an et les emplois intérimaires dont le contrat débute en 2019.

[37] Sont croisés ici les nombres d'emplois ayant débuté depuis moins d'un an dans l'enquête Emploi déclinés dans la nomenclature des activités en 21 niveaux (NAF 21) et dans la nomenclature des familles professionnelles en 22 niveaux (FAP 22). Cette matrice est utilisée pour obtenir la déclinaison des emplois vacants et des DPAE (originellement en secteurs d'activité) par famille professionnelle (FAP 22). L'annexe 1 permet d'apprécier l'effet de l'utilisation de cette matrice à partir d'un test sur les offres collectées par Pôle emploi 2019.

TABLEAU 6 – Décomposition par domaine professionnel de la dissimilarité (Duncan) entre les offres en ligne (Jocas) et des sources de référence

En %

Domaine professioinel (FAP 22)	Offres STMT et partenaires	Emplois de moins d'un an (EEC)	Projets de recrutement (BMO)	Emplois vacants (Acemo)*	Intentions d'embauches (DPAE)*	Embauches (MMO)*
Bâtiment, travaux publics	0,9	1,9	2,2	1,5	1,5	3,4
Maintenance	0,5	2,3	2,3	2,0	2,4	2,4
Ingénieurs et cadres de l'industrie	0,7	1,0	1,3	0,9	1,3	1,2
Gestion, administration des entreprises	0,7	2,1	4,3	1,6	2,9	2,6
Informatique et télécommunications	1,4	1,8	2,1	1,1	2,4	1,4
Études et recherche	0,5	0,3	0,5	0,1	0,6	0,4
Commerce	0,4	0,7	2,2	0,2	3,2	2,2
Agriculture, marine, pêche	0,4	1,0	5,1	0,0	0,6	0,1
Électricité, électronique	0,1	0,0	0,0	0,1	0,2	0,1
Industries de process	0,1	0,3	0,2	0,3	1,3	0,1
Matériaux souples, bois, industries graphiques	0,2	0,2	0,2	0,3	0,3	0,1
Transports, logistique et tourisme	0,3	1,1	0,3	0,6	3,6	0,3
Artisanat	0,0	0,2	0,0	0,2	0,2	0,2
Hôtellerie, restauration, alimentation	1,0	1,0	2,9	2,4	1,5	3,4
Services aux particuliers et aux collectivités	2,1	2,3	3,1	1,0	3,9	3,2
Communication, information, art et spectacle	0,1	0,7	1,6	1,1	0,8	2,0
Santé, action sociale, culturelle et sportive	0,3	2,0	2,7	1,2	1,1	4,4
Mécanique, travail des métaux	0,0	0,4	0,9	0,4	0,6	0,9
Administration publique, professions juridiques, armée et police	0,1	1,0	0,3	0,4	0,8	1,1
Banque et assurances	0,0	0,4	0,1	0,3	0,3	0,3
Enseignement, formation	0,6	0,2	0,1	0,2	0,2	0,2
Dissimilarité de Duncan	**10,4**	**20,9**	**32,2**	**15,9**	**29,4**	**29,9**

* Sources non déclinables par métier ; une matrice de passage secteur – métier a été utilisée pour convertir les secteurs NAF 21 en domaines professionnels FAP 22 (Annexe 1).

Lecture : 32,2 % des offres scrapées en ligne (Jocas) devraient changer de domaine professionnel pour que la répartition de leurs domaines s'aligne sur celle des projets de recrutement issus de l'enquête BMO. La contribution la plus importante à cette dissimilarité est due au domaine de l'agriculture, de la marine et de la pêche (5,1 points sur 32,2). Ce domaine compte relativement plus de projets de recrutement que d'offres scrapées (en rouge), contrairement au domaine du bâtiment (en jaune).

Champs : France (hors Mayotte) et France métropolitaine pour MMO, 2019.

Source : Jocas, STMT (Dares-Pôle emploi), offres partenaires de Pôle emploi, BMO (Pôle emploi), EEC (Insee), DPAE (Urssaf), Acemo (Dares) et MMO (Dares) - calculs Dares.

En décomposant la dissimilarité entre la base Jocas 2019 et les autres sources de référence, trois groupes de domaines peuvent être distingués en fonction de leur contribution à l'indice de dissimilarité : les domaines surreprésentés (en jaune), les sous-représentés (en rouge), et ceux – plutôt bien représentés – pour lesquels ne se dégage pas une tendance nette pour toutes les sources.

Le premier groupe est composé des domaines qui sont systématiquement surreprésentés par Jocas (en jaune), quelle que soit la source de référence considérée. Ce groupe est composé du domaine du bâtiment et des travaux publics, de deux domaines de l'industrie (maintenance, ingénieurs et cadres de l'industrie) et de quatre domaines du tertiaire (gestion et administration des entreprises, informatique et télécommunications, études et recherche, commerce).

La surreprésentation par Jocas de ces domaines tient à plusieurs facteurs. En premier lieu, ce sont des domaines professionnels qui effectuent beaucoup de recrutements en ligne : d'après l'enquête Ofer, à l'exception du domaine du bâtiment et des travaux publics, les employeurs dans ces domaines utilisent plus souvent le canal « Internet » pour recruter. C'est notamment le cas

dans l'informatique et les télécommunications, où une offre d'emploi est diffusée sur Internet lors de deux recrutements sur trois[38] (Tableau 7). Pour les autres domaines, le dépôt d'une offre en ligne est utilisé dans au moins la moitié des recrutements.

TABLEAU 7 - Caractéristiques des offres par domaines professionnels

Domaine professionnel (FAP 22)	Part des Recrutements avec annonce en ligne	Part des cadres dans l'emploi	Part des recrutements multiples
Bâtiment, travaux publics	29	10	28
Maintenance	53	0	26
Ingénieurs et cadres de l'industrie	54	100	27
Gestion, administration des entreprises	54	28	20
Informatique et télécommunications	66	66	45
Études et recherche	54	100	37
Commerce	51	23	32
Agriculture, marine, pêche	44	4	50
Électricité, électronique	35	0	29
Industries de process	40	0	30
Matériaux souples, bois, industries graphiques	54	0	41
Transports, logistique et tourisme	51	5	41
Artisanat	41	0	38
Hôtellerie, restauration, alimentation	45	5	30
Services aux particuliers et aux collectivités	55	0	39
Communication, information, art et spectacle	48	62	18
Santé, action sociale, culturelle et sportive	51	10	40
Mécanique, travail des métaux	30	0	19
Administration publique, professions juridiques, armée et police	58	25	24
Banque et assurances	64	37	47
Enseignement, formation	39	40	34
Ensemble	**49**	**18**	**32**

Note : domaines surrepésentés (jaune) ; domaines sous-représentés (rouge) ; autres domaines (blanc).

Lecture : en 2016, 66 % des recrutements dans le domaine de l'informatique et des télécommunications ont eu recours à une diffusion d'offres d'emploi en ligne. 66 % des emplois dans l'informatique et les télécommunications sont sur des postes de cadres. 45 % des recrutements sont des recrutements multiples.

Champs : Ofer : ensemble des nouveaux recrutements en CDI ou en CDD de plus d'un mois entre septembre et novembre 2015 des établissements d'au moins un salarié du secteur concurrentiel à l'exception du domaine professionnel politique, religion ; France. Enquête emploi : salariés en emploi en France.

Source : enquête Ofer 2016, Dares et enquête Emploi 2019, Insee.

[38] Un agument très souvent avancé est aussi que les recruteurs de ce domaine dynamique avec un fort turnover diffusent parfois des offres dans le but de se constituer une base de données de CV pour leurs futurs recrutements.

Ensuite, la base Jocas comptabilise, à ce stade, une seule fois chaque offre même si elle vise plusieurs recrutements du même type. Dans la base, les domaines qui ont plus souvent recours à des recrutements multiples sont donc sous-représentés et par conséquent, les autres surreprésentés. Cela explique la surreprésentation du domaine de la gestion et l'administration des entreprises. Il compte seulement 20 % des recrutements multiples (Tableau 7), ce qui est faible comparé à la moyenne tous domaines confondus qui est de 32 %. De même, le domaine du bâtiment et des travaux publics ainsi que ceux de la maintenance et des ingénieurs et cadres de l'industrie ont aussi une part de recrutements multiples inférieure à la moyenne.

Une troisième explication possible est liée à la part des cadres dans ces domaines. Les domaines surreprésentés par Jocas sont des domaines avec une forte présence de cadres[39]. Pour certains d'entre eux, comme les ingénieurs et cadres de l'industrie et le domaine des études et de la recherche, ils sont composés exclusivement de cadres, alors que dans d'autres domaines, les cadres se situent dans certaines de leurs familles professionnelles. Pour ces derniers, l'analyse du Duncan au niveau de leurs familles professionnelles (Annexe 3) montre que la dissimilarité observée au niveau de ces domaines est captée par les familles de cadres. Ainsi, dans le bâtiment et les travaux publics, c'est la famille des cadres du bâtiment et des travaux publics qui contribue le plus à la dissimilarité. Dans les domaines de la gestion et de l'administration des entreprises, de l'informatique et des télécommunications ainsi que du commerce, une grande part de la dissimilarité est aussi liée aux familles de cadres (respectivement des services administratifs, comptables et financiers, des ingénieurs de l'informatique et des cadres commerciaux et technico-commerciaux).

Le second groupe est composé de domaines professionnels sous-représentés systématiquement par Jocas (en rouge dans le Tableau 6) ; on y retrouve des domaines du tertiaire de l'agriculture[40] et dans une moindre mesure certains domaines industriels. Les domaines industriels concernés sont l'électricité et l'électronique, les industries de process[41] et le domaine des matériaux souples, bois et industries graphiques. Dans le tertiaire, figurent les domaines suivants : transport, logistique et tourisme ; artisanat ; hôtellerie, restauration et alimentation ; services aux particuliers et aux collectivités ; communication, information et art et spectacle ; santé, action sociale, culturelle et sportive.

Les éléments d'explications avancés pour le premier groupe peuvent intervenir ici, dans un sens opposé. Ainsi, pour plusieurs de ces domaines, le recours à la diffusion d'offres d'emploi en ligne lors du recrutement est beaucoup moins important pour des métiers plutôt manuels, d'aides à la personne et de contact avec un public (LHOMMEAU et REMY, 2021). Il est alors logique que Jocas les sous-représente. Dans l'hôtellerie et la restauration par exemple, la plupart des recrutements (récurrents) se font sur un marché secondaire externe ; la diffusion d'offres en ligne y est plus rare, d'autres canaux de recrutement étant privilégiés, comme les relations, les intermédiaires publics ou les candidatures spontanées. Ainsi, seulement 45 % des recrutements en CDI et CDD de plus d'un mois donnent lieu à la diffusion d'une offre d'emploi en ligne (Tableau 7) et très souvent auprès d'un intermédiaire public (FONDEUR, 2013 ; LHOMMEAU et REMY, 2021). Pour les

[39] À l'exception du domaine de la maintenance, qui est le seul domaine ne comptant pas de cadre mais des techniciens et agents de maitrise et des ouvriers.

[40] À l'exception des emplois vacants d'Acemo par rapport auxquels Jocas n'est ni sous-représenté ni surreprésenté. La principale raison est que le secteur de l'agriculture n'est pas pris en compte dans l'enquête Acemo. Nous ajoutons donc l'agriculture au groupe des domaines sous-représentés par Jocas.

[41] Ces deux premiers domaines sont légèrement surreprésentés par la base Jocas par rapport à la source MMO. Ils sont inclus dans ce second groupe, car il s'agit de la seule source dans cette situation et les contributions sont faibles.

services aux particuliers et aux collectivités, le recours à une offre en ligne (55 % des recrutements) est plutôt élevé mais il est lié à une forte intermédiation de sites spécialistes (Tableau 7).

La sous-représentation peut aussi provenir de la faible présence voire l'absence de cadres dans ces domaines. En effet, la moitié de ces métiers ne comptent aucun cadre parmi leurs effectifs, et, dans les autres, la part des cadres du domaine varie entre 4 % dans l'agriculture, 5 % dans le transport, la logistique et le tourisme ainsi que dans l'hôtellerie, la restauration et l'alimentation et 10 % dans la santé, l'action sociale, culturelle et sportive (Tableau 7). Le domaine de la communication, de l'information et de l'art et du spectacle fait exception, avec près de deux tiers de cadres. Or, l'effort de recrutement est plus important pour le recrutement de cadres, qui concernent très souvent des postes durables et avec de bonnes conditions d'emploi ; les recruteurs sont incités à faire plus d'efforts et notamment à diffuser leurs offres en ligne (FONDEUR, 2013). Ainsi, puisque les domaines de ce groupe comptent très peu, voire pas du tout de cadres dans leurs effectifs (Tableau 7), la diffusion d'offres d'emploi en ligne est moins souvent privilégiée, d'où la sous-représentation de ces domaines par Jocas.

Dans ces domaines, les offres d'emploi portent très souvent sur des recrutements multiples. Jocas les sous-représente donc, en comparaison aux autres domaines où ce phénomène est moins fréquent. C'est notamment le cas de l'agriculture, où un recrutement sur deux est multiple (contre moins d'un sur trois en moyenne, Tableau 7), ce qui en fait le domaine qui compte la proportion la plus élevée de recrutements multiples. D'autres domaines de ce groupe sont également très concernés par les recrutements multiples : le transport, la logistique et le tourisme (41 %) ; les services aux particuliers et aux collectivités (39 %) ; les matériaux souples, le bois et les industries graphiques ; etc.

Le dernier groupe rassemble les domaines professionnels où il ne se dégage pas une caractérisation nette (sur- ou sous-représentation systématique quelle que soit la source de référence considérée). Par exemple, le domaine de l'enseignement et de la formation est plutôt sous-représenté par Jocas par rapport aux offres diffusées par Pôle emploi et aux embauches MMO, mais il apparaît légèrement surreprésenté par rapport aux autres sources. Cette absence de caractérisation nette s'observe également pour le domaine de la banque et des assurances, ou encore certains domaines de l'industrie, comme la mécanique et le travail des métaux.

Globalement, les contributions à la dissimilarité dans ce dernier groupe sont faibles et semblent davantage liées aux caractéristiques de chacune des sources références. Par exemple, la sous-représentation de Jocas par rapport aux DPAE, aux emplois vacants et aux embauches (MMO) dans le domaine de l'administration publique, les professions juridiques, l'armée et la police pourrait être liée à un problème de précision de la matrice de passage utilisée pour passer du secteur au domaine professionnel pour ces sources. Cependant, cet argument est moins valable pour les autres domaines du troisième groupe car le sens de la dissimilarité avec Jocas n'est pas le même pour les trois sources qui dépendent toutes de la matrice de passage. La non-prise en compte de l'emploi intérimaire ainsi que de l'absence des secteurs agricoles et des emplois publics dans les emplois vacants peuvent contribuer à ces écarts. La non couverture de certains secteurs par l'enquête Acemo et MMO impacte différemment les domaines. Le domaine agricole sera par exemple plus affecté que les autres par l'absence du secteur agricole. Selon la matrice de passage, le secteur agricole compte 84 % des travailleurs dans le domaine professionnel de l'agriculture, contre 0 à 4 % pour les autres domaines.

Couverture du territoire

En ce qui concerne la couverture du territoire national, le Tableau 8 ci-dessous donne une décomposition de l'indicateur de dissimilarité de Duncan entre la structure régionale des offres de quatre des cinq sources[42] (hors Acemo) et celle des offres Jocas. Les valeurs de l'indicateur sont faibles, surtout avec les offres diffusées par Pôle emploi (y compris offres partenaires). Cela signifie que les offres Jocas ont une couverture du territoire assez similaire à celles des autres sources de référence. Seulement 2 % des offres Jocas devraient être redistribuées géographiquement pour s'aligner sur la distribution régionale des offres Pôle emploi. Ce chiffre monte à 8, 9 et 11 % avec les projets de recrutements BMO, les DPAE et les embauches MMO, respectivement, et à 5 % avec les nouveaux emplois de moins d'un an de l'enquête Emploi. La région contribuant le plus à la dissimilarité est l'Île-de-France, où les offres Jocas sont fortement surreprésentées notamment par rapport aux projets BMO et aux DPAE, mais aussi très fortement sous-représentées par rapport aux embauches MMO. Les offres Jocas sont également surreprésentées - toutes sources confondues - dans la région Auvergne Rhône-Alpes. Une plus grande concentration des cadres dans ces deux régions est probablement à l'origine de cette plus importante dissimilarité.

TABLEAU 8 - Décomposition par région de la dissimilarité (Duncan) entre les offres en ligne (Jocas) et des sources de référence

En %

Région	Offres STMT et partenaires	Emplois de moins d'un an (EEC)	Projets de recrutement (BMO)	Intentions d'embauches (DPAE)	Embauches (MMO)
Île-de-France	0,0	0,3	1,7	2,3	4,34
Centre-Val de Loire	0,2	0,3	0,5	0,0	0,65
Bourgogne-Franche-Comté	0,1	0,1	0,1	0,3	0,34
Normandie	0,1	0,0	0,0	0,7	0,08
Hauts-de-France	0,0	0,9	0,8	1,7	0,34
Grand Est	0,1	0,6	0,1	0,4	0,87
Pays de la Loire	0,1	0,6	0,3	0,2	0,93
Bretagne	0,1	0,0	0,2	0,3	0,40
Nouvelle-Aquitaine	0,0	0,1	0,4	0,3	0,45
Occitanie	0,3	0,7	1,0	0,5	0,22
Auvergne-Rhône-Alpes	0,3	0,8	1,0	1,6	1,79
Provence-Alpes-Côte d'Azur	0,2	0,3	1,3	0,1	0,55
Corse	0,0	0,1	0,3	0,0	0,06
Dissimilarité de Duncan	**1,6**	**4,6**	**7,8**	**8,5**	**11,0**

Lecture : 8,5 % des offres en ligne (Jocas) (dont 2,3 % en Île-de-France) devraient changer de région pour que leur répartition par région s'aligne sur celle des intentions d'embauches (DPAE). Dans Jocas, relativement aux DPAE, les offres collectées en ligne sont surreprésentées en Île-de-France (en jaune) alors qu'elles sont sous-représentées dans les Hauts-de-France (en rouge).

Champs : France (hors Mayotte) et France métropolitaine pour MMO, 2019.

Source : Jocas, STMT, offres partenaires, BMO (Pôle emploi), EEC (Insee), DPAE (Urssaf) et MMO (Dares) - calculs Dares

[42] Les enquêtes Acemo ne sont pas représentatives géographiquement. De plus, leurs pondérations et leurs résultats sont exprimés au niveau des entreprises (et non des établissements) et elles ne peuvent donc pas être exploitées pour des problématiques locales.

Ce cadrage montre que la base Jocas offre une information complémentaire aux autres sources d'information existantes, bien qu'elle surreprésente certains domaines professionnels ou territoires.

Pour mieux estimer sa représentativité, il faudrait disposer d'éléments de calage, afin de redresser la base Jocas. Une enquête Ofer « élargie » et plus régulière pourrait servir de référence pour les offres en ligne : elle inclurait les recrutements n'ayant pas abouti et serait étendue à tous les types de contrats (alors qu'elle est restreinte aux CDI et CDD de plus d'un mois dans sa dernière édition).

3. Quelques usages de la base Jocas

Comme toujours en statistique, la représentativité d'une source et l'incertitude qui entoure les résultats qui en sont issus doivent être considérées au regard de l'usage qui en est fait. Ainsi, la représentativité des données Jocas est considérée comme acceptable pour certains usages, que présente cette dernière section : indicateurs ayant déjà fait l'objet d'une publication, travaux en cours ou encore idées d'exploitations de la base Jocas.

a. Deux indicateurs déjà publiés : les tensions sur le marché du travail et un suivi hebdomadaire des offres en ligne durant la crise du Covid-19

À ce jour, deux utilisations de la base Jocas ont fait l'objet de publications par la Dares. Tout d'abord, la base Jocas a été utilisée pour le calcul des tensions sur le marché du travail (NIANG et VROYLANDT, 2020). Historiquement, les données d'offres d'emploi utilisées pour le calcul des tensions sont les offres d'emploi déposées à Pôle emploi (STMT). Or, ces données ne sont pas représentatives de l'ensemble des recrutements ni des offres en ligne (cf. partie I.2), notamment pour les métiers les plus qualifiés. De plus, les offres de Pôle emploi sont soumises à des variations dues à des modifications de l'écosystème du marché du travail numérique. Afin de compléter les offres de la STMT mais aussi de rendre le champ plus robuste, les offres Jocas ont donc été agrégées aux offres déposées à Pôle emploi pour le calcul des tensions. En 2019, l'ajout des offres Jocas aux offres de la STMT a eu pour effet d'accroître la mesure de la tension sur les métiers d'ingénieurs et de cadres et de la réduire sur les métiers d'employés et d'ouvriers (Eurostat, 2020). En effet, bien que la tension[43] ait été faiblement modifiée pour la majorité des métiers (moins de 0,1 de différence en valeur absolue entre le calcul avec et sans Jocas), elle a nettement augmenté pour certains domaines professionnels très qualifiés. C'est notamment le cas des ingénieurs et des cadres de l'industrie (+0,51 point en moyenne sur l'ensemble des métiers du domaine), de l'informatique et des télécommunications (+0,30), et du domaine des études et de la recherche (+0,30). Au contraire, la tension a fortement baissé pour le domaine de l'électricité et de l'électronique (-0,37).

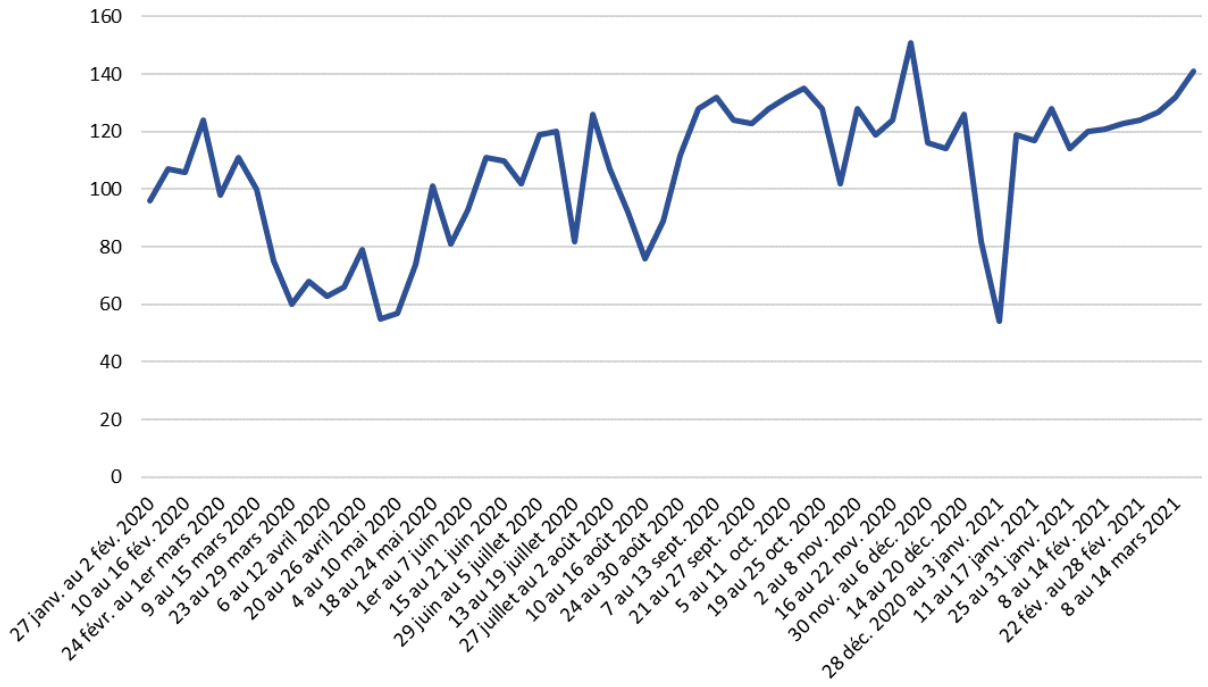
Ensuite, les offres Jocas ont fait l'objet en 2020 et 2021 d'une publication dans le cadre du suivi du marché du travail durant la crise du Covid-19 (Figure 8). Les données Jocas permettent de donner une visibilité hebdomadaire de l'évolution du flux d'offres postées en ligne, et de mieux apprécier l'impact des mesures sanitaires sur le marché du travail en France. Ces données « hautes fréquences » sont particulièrement adaptées pour donner une première estimation en temps quasi-réel d'un choc économique de grande ampleur : contrairement aux enquêtes ou aux données administratives, les données Jocas sont relevées le lendemain de leur parution. Cette publication est par ailleurs complétée par un commentaire intégrant des éléments qualitatifs sur les domaines professionnels et les types de contrat. Cependant, si cet indicateur donne une

[43] Nous faisons ici référence à l'indicateur synthétique de tension, qui est centré et varie entre -3 et 3.

tendance globale sur l'offre d'emploi en France, il est parfois difficile d'interpréter ses variations au-delà du très court terme. En effet, la part des offres couverte par un site peut évoluer rapidement, ce qui perturbe les analyses sur plusieurs mois.

Des publications de ce genre (plus particulièrement durant la crise sanitaire du Covid-19) existent aussi dans d'autres pays comme le Royaume-Uni ou l'Allemagne. Au Royaume-Uni, l'ONS (office national des statistiques) utilise les données du site d'offres d'emploi en ligne « Adzuna » pour produire un indicateur de suivi conjoncturel du marché du travail par catégorie d'emploi comparable au métier. En Allemagne, Destatis utilise les offres d'emploi en provenance de cinq sites (y compris Adzuna) pour suivre l'évolution du marché du travail durant la crise du Covid-19. Les évolutions observées dans ces deux pays à partir de ces données sont semblables à celles de la base Jocas[44].

FIGURE 8 - Évolution hebdomadaire du flux d'offres d'emploi postées en ligne, au 21 mars 2021



Note : indice base 100 lors de la semaine du 9 au 15 mars 2020. Les données des semaines du 28 septembre au 18 octobre 2020 ont été ajustées pour ne pas répercuter l'évolution forte, temporaire et inexpliquée de l'un des sites.
Source : Panel de 12 sites d'offres d'emploi, calcul Dares.

[44] Les estimations d'offres d'emploi diffusées en ligne (« Online job advert estimates ») au Royaume-Uni : https://www.ons.gov.uk/economy/economicoutputandproductivity/output/datasets/onlinejobadvertestimates ; L'indice d'offres d'emploi en ligne (« online job index ») en Allemagne : https://www.destatis.de/EN/Service/EXDAT/Datensaetze/online-job-index.html

b. De nouvelles exploitations, déjà initiées ou à venir

Suite à l'arrêt de la publication hebdomadaire des offres en ligne en juin 2021, un indicateur trimestriel du niveau mensuel d'offres par métier verra le jour en 2022. Il prendra notamment en compte les variations de champ (ajout ou suppression de sites Internet ; prise importante de part de marché par un site). A l'avenir, cet indicateur pourra être consolidé, avec par exemple l'amélioration de la couverture des sites d'emploi en ligne ou la correction des jours ouvrés.

Les données Jocas peuvent aussi nous permettre de mieux comprendre le marché du travail numérique, en analysant les comportements des recruteurs sur les sites d'offres en ligne et les interactions entre les différents sites. Par exemple, des travaux exploratoires sont engagés sur l'analyse de la durée pendant laquelle un recruteur laisse une offre en ligne : une durée de publication plus longue pourrait être corrélée à de plus grandes difficultés de recrutement. De même, la volonté des recruteurs d'augmenter la visibilité de leur offre (qu'il serait possible d'évaluer *via* la longueur de la description ou le nombre de doublons par offre) pourrait signaler un manque de candidats.

Cependant, les analyses quantitatives se heurtent pour l'instant à la difficulté d'évaluer la représentativité des données Jocas (cf. partie III.2). De même, les analyses en évolution sont limitées par de possibles instabilités du champ dues à l'évolution des parts de marché des sites scrapés mais aussi aux changements de technologie ou d'offre commerciale opérés par les sites, qui peuvent inciter les recruteurs à modifier leurs comportements.

De manière plus qualitative, ce flux continu de données non structurées permet d'avoir accès à des informations habituellement lissées par le format contraint des enquêtes ou des sources administratives. Par exemple, la base Jocas peut renseigner sur l'évolution des intitulés des métiers et des compétences et ainsi contribuer à la mise à jour des nomenclatures. Cela est particulièrement vrai dans des domaines comme le numérique, où métiers et compétences (comme les langages de programmation) évoluent rapidement (DESJONQUERES et *al*., 2019). La base Jocas, par sa grande taille, peut également servir de « base de sondage » pour du « testing » ou des enquêtes qui nécessiteraient d'entrer en contact avec des recruteurs, sur un métier donné par exemple.

Enfin, si ce document ne détaille que deux aspects de la base Jocas 2019 (métier et lieu d'activité), une restructuration d'un plus grand nombre de champs présents dans les offres en ligne est en cours. Les travaux de recodification portent notamment sur le type et la durée du contrat, le salaire et le secteur. Des analyses plus complexes sont également prévues, comme l'extraction des compétences requises dans le texte des offres et la codification de l'établissement employeur (SIRET) afin d'apparier les offres à des bases de données administratives telles que les DPAE ou la déclaration sociale nominative (DSN)[45].

[45] Ces déclarations de paies réalisées chaque mois par les employeurs permettent de connaître le détail des caractéristiques associées aux contrats de travail, telles que l'établissement employeur, le secteur d'activité, la nature du contrat, etc.

Conclusion

Ce document de travail porte sur la base de données Jocas, constituée des offres d'emploi en ligne que la Dares collecte depuis 2018 auprès d'un échantillon d'une quinzaine de sites d'offres d'emploi en ligne. Il revient sur les différentes étapes du processus de collecte et de traitement de ces offres jusqu'à leur déclinaison dans la nomenclature des familles professionnelles (FAP). Il confronte ensuite ces offres d'emploi en ligne aux sources usuelles sur le marché du travail, afin d'apprécier la valeur de cette nouvelle source de données.

La base de données Jocas a été construite en utilisant un mode de collecte innovant : le *webscraping*. Ce document justifie l'utilisation de cette méthode et décrit sa mise en œuvre effective : développement de scripts Python, utilisation de machines virtuelles, respect d'une « nétiquette » afin de ne pas perturber l'activité des sites cibles. Au-delà de la collecte, ce document décrit le traitement des offres et notamment leur classification par métier (*via* le code ROME). Là encore, une méthode innovante pour la Dares a été utilisée : un algorithme de *machine learning* entraîné sur des offres annotées. Les performances obtenues sont satisfaisantes (F1-score de 0,96 au niveau le moins détaillé de la nomenclature), quoiqu'inégales selon les métiers. Ensuite, les offres d'emploi ont été dédupliquées, et enfin transposées en FAP, nomenclature utilisée à la Dares. Malgré les difficultés rencontrées, le *scraping* et les étapes ultérieures de traitement ont finalement permis de constituer la base Jocas, composée d'un peu plus de 5 millions d'offres d'emploi en 2019.

Si l'enquête Offre d'emploi et recrutement (Ofer) indique que les offres publiées sur Internet ne représentent qu'une partie (biaisée) des recrutements (49 % en 2016), il est à l'inverse difficile d'estimer la représentativité d'une source d'offres en ligne. La base Jocas se heurte donc à des problèmes - non résolus - de représentativité. À ce jour, aucune source administrative ou d'enquête ne permet d'estimer précisément la couverture de Jocas.

À défaut, ce document fournit des premiers cadrages, en comparant Jocas avec les sources usuelles exploitées pour effectuer le suivi du marché du travail. Cette analyse a été effectuée par rapport aux offres diffusées par Pôle emploi, aux projets de recrutement estimés par l'enquête besoins en main-d'œuvre (BMO), aux emplois de moins d'un an issus de l'enquête Emploi, aux déclarations d'embauches, aux emplois vacants et aux embauches MMO, sur la base de leur couverture des métiers et du territoire.

Concernant la couverture des métiers, Jocas se rapproche des offres diffusées en ligne par Pôle emploi. Avec les autres sources, la dissimilarité de Duncan – mesurée au niveau FAP 22 – est nettement plus élevée et atteint jusqu'à 32 % avec BMO, source dont Jocas s'éloigne le plus. Les différences de représentativité peuvent s'expliquer par différents facteurs. Les domaines professionnels effectuant beaucoup de recrutements en ligne (par exemple le domaine de l'informatique et des télécommunications) ou ayant une forte part de cadres (tels que les ingénieurs et cadres de l'industrie ou le domaine des études et de la recherche) sont plutôt surreprésentés dans Jocas. À l'inverse, les domaines moins qualifiés, comptabilisant une grande proportion de recrutements multiples (notamment l'agriculture) ou mobilisant des canaux de recrutement informels (comme l'hôtellerie et la restauration) ont tendance à être sous-représentés par Jocas.

Concernant l'aspect territorial, Jocas présente une couverture régionale du territoire comparable aux sources de référence. Certaines régions comme l'Île-de-France et Auvergne Rhône-Alpes sont toutefois surreprésentées, en lien avec la surreprésentation des métiers de cadres dans ces deux régions.

La source Jocas a déjà fait l'objet d'exploitations dans la production de statistiques. Elle a ainsi servi dans l'analyse des tensions sur le marché du travail. En effet, l'une des améliorations du

nouvel indicateur de tension a été la prise en compte des offres Jocas (en plus des offres collectées par Pôle emploi), afin de mieux couvrir les offres d'emploi sur le marché du travail, notamment pour certains métiers de cadres. Les offres scrapées ont également été utilisées pour la production du tableau de suivi de la situation du marché du travail en 2020-2021 lors de la crise du Covid-19 et un indicateur de l'évolution mensuelle du nombre d'offres par métier verra le jour en 2022.

D'autres exploitations des données Jocas sont en cours ou à venir dans le but toujours de mieux comprendre le marché du travail, plus particulièrement celui numérique. Ces exploitations portent sur des domaines variés, notamment l'analyse les comportements des recruteurs sur les sites d'offres en ligne et les interactions entre les différents sites.

Ce document vise donc une première présentation de Jocas, cette nouvelle source sur le marché du travail (numérique) qui est encore en développement. Elle va continuer d'évoluer dans son processus d'élaboration (amélioration du codage par métier et de la déduplication, enrichissement des données collectées…) afin d'en fiabiliser et d'accroître ses apports statistiques.

Dans cette perspective, il serait intéressant de mettre en place une enquête plus régulière sur le recrutement en ligne, afin de mieux cerner la représentativité des offres Jocas. Une telle enquête pourrait se faire dans le cadre d'une nouvelle édition de l'enquête Ofer, qui ne porterait plus uniquement sur les recrutements ayant abouti, mais sur tous les recrutements.

Bibliographie

AVENTUR F., BONNET A., DE VISME N. (2017), « La place du numérique dans la recherche de candidats par les employeurs ? », *Éclairages et Synthèses,* n° 29, Pôle emploi.

AZAR J., MARINESCU I., STEINBAUM M., TASKA B. (2020), « Concentration in US labor markets: Evidence from online vacancy data », *Labour Economics 66(101886).*

BANFI S., CHOI S., VILLENA-ROLDÁN B. (2019), « Sorting On-line and On-time », *Bristol Economics Discussion Papers, 19/706*.

BERGEAT M., REMY V. (2017), « Comment les employeurs recrutent-ils leurs salariés ? », *Dares Analyses*, n° 64.

BERGEAT M., MINNI C., REMY V., FONDEUR Y. (2018), « Mobiliser Internet pour recruter : quelles sont les pratiques des employeurs ? », *Dares Analyses,* n° 32.

BESSY C., MARCHAL E. (2006) , « La mobilisation d'Internet pour recruter : aux limites de la sélection à distance », *La Revue de l'Ires*, 2006/3, n° 52, p. 11-39.

BOSELLI R., CESARINI M., MERCORIO F., MEZZANZANICA M. (2018), « Classifying online job advertisements through machine learning », *Future Generation Computer Systems,* vol. 86, p. 319-328.

Cedefop (2019). Online job vacancies and skills analysis: a Cedefop pan-European approach. Luxembourg: Publications Office. http://data.europa.eu/doi/10.2801/097022

Conseil d'orientation pour l'emploi (2013), « Emplois durablement vacants et difficultés de recrutement », *Rapport technique du COE*.

Conseil d'orientation pour l'emploi (2015), « L'impact d'internet sur le fonctionnement du marché du travail », *Rapport technique* du COE.

COUSTEAUX A.-S. (2019), « Des ménages et des entreprises de plus en plus connectés, mais des disparités persistantes », L'économie et la société à l'ère du numérique, *Insee Références*.

DAVIS S., SAMANIEGO DE LA PARRA B. (2017), « Application Flows », *Document de travail*.

DESJONQUERES A., DE MARICOURT C., MICHEL C. (2019) « Data scientists, community managers… et informaticiens : quels sont les métiers du numérique ? », L'économie et la société à l'ère du numérique, *Insee Références*, édition 2019.

Eurostat (2020), « ESSnet Big Data II, WPB report on the statistical output, required quality and definition of the necessary metadata at European and national level ».

FABERMAN J., KUDLYAK M. (2016), « What does online job search tell us about the labor market? », *Economic perspectives*, *40*(1), p. 1-15.

FONDEUR Y., LHERMITE F. (2013), « Outils informatiques de gestion de recrutement et standardisation des façons de recruter », *Document de travail du CEET*, n° 165.

FONDEUR, Y. (2013). « Introduction. Systèmes d'emploi et pratiques de recrutement », *La Revue de l'IRES*, vol. 76, n°1, p. 31-43.

FONDEUR Y., LHERMITTE F. (2006), « Réseaux sociaux numériques et marché du travail », *La Revue de l'IRES*, 2006/3, n° 52, p. 101-131.

FONDEUR Y. (2016), « Dynamiques écologiques du marché du travail en ligne autour de la circulation des offres d'emploi », *Études et recherches*, n° 7, Pôle emploi.

JACKSON M. (2007), « How far merit selection? Social stratification and the labour market », *The British journal of sociology*, *58*(3), p. 367-390.

KUREKOVÁ L. M., BEBLAVY M., THUM A. E. (2014), « Using internet data to analyse the labour market: a methodological enquiry », *IZA Discussion Papers*, n° 8555.

LENAERTS K., BEBLAVÝ M., FABO B. (2016), « Prospects for utilisation of non-vacancy Internet data in labour market analysis—an overview », *IZA Journal of Labor Economics 5*(1).

LHOMMEAU B., REMY V. (2021), « Les critères de sélection : un résumé du processus du recrutement selon le métier », Dares Document d'études.

LITTLE R. J. A., RUBIN D. B. (2002), « Statistical Analysis with Missing Data, 2nd Edition », New York: John Wiley & Sons.

LONGFORD N. T. (2005), « Missing data and small-area estimation: Modern analytical equipment for the survey statistician », Springer Science & Business Media.

DE MARICOURT C. (2018), « Collecte et analyse d'offres d'emploi en ligne », *PFE - Projet de fin d'études*, ENSTA.

MARTIKAINEN J. (2010), « Weighting and estimation methods: JVS estimation in Finland by Horowitz-Thomson-Type estimator », *1st and 2nd international workshops on methodologies for job vacancy statistics*, Eurostat.

NIANG M., VROYLANDT T. (2020), « Les tensions sur le marché du travail en 2019 », *Dares Résultats*, n° 032, Dares.

ATILF (2019), « Morphalou [Lexique] », ORTOLANG (Open Resources and TOols for LANGuage).

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., VANDERPLAS J. (2011), « Scikit-learn: Machine learning in Python », *The Journal of machine Learning research 12*, p.2825-2830.

ROYALL R. M. (1992), « The model based (prediction) approach to finite population sampling theory », *Lecture Notes-Monograph Series*, vol. 17, p.225-240.

SINCLAIR T., GIMBEL M. (2020), « Mismatch in Online Job Search », *IIEP working paper*, n° 2020-1.

ŠTEFÁNIK M. (2012), « Internet job search data as a possible source of information on skills demand (with results for Slovak university graduates) », *Building on skills forecasts—Comparing methods and applications*, 246.

VAN OURS J., RIDDER G. (1992), « Vacancies and the recruitment of new employees », *Journal of Labor Economics*, *10*(2), p.138-155.

Annexes

ANNEXE 1 - Test de la matrice de passage secteurs-métiers sur les offres collectées par Pôle emploi en 2019

Deux des sources de comparaison – les emplois vacants issus des enquêtes Acemo et les déclarations préalables à l'embauche (DPAE) - ne sont pas déclinées par métier mais par secteur d'activité (NAF). Par ailleurs, la source Jocas n'est pas disponible par secteur d'activité mais seulement par métier (FAP). Afin de comparer Jocas avec Acemo et les DPAE, il est nécessaire de décliner par métier les données figurant dans ces deux dernières sources. La méthode choisie consiste à créer une matrice de passage secteurs-métiers. La matrice de passage a été établie à partir des nombres d'emploi depuis moins d'un an issus de l'enquête Emploi en continu sur la période 2015-2019. Elle permet de convertir une source de la NAF 21 vers la FAP 22.

Pour évaluer la performance de cette matrice, elle est appliquée à une source déclinable à la fois par secteur et par métier : les offres déposées à Pôle emploi en 2019, issues de la statistique du marché du travail (STMT). La comparaison entre la distribution prédite et celle réellement observée permet d'estimer le biais induit par l'utilisation de la matrice de passage. En l'occurrence, l'indice de dissimilarité de Duncan entre la distribution par métier prédite et celle observée est de 11 % (Tableau A1). Au-delà de cette mesure globale, la déclinaison des différences par domaine professionnel permet d'identifier lors des analyses les écarts imputables à l'utilisation de la matrice de passage.

TABLEAU A1 - Dissimilarité par domaines professionnels entre la distribution observée et celle obtenue avec l'application de la matrice secteurs-métiers

En %

Domaine professionnel (FAP 22)	Offres STMT 2019
Agriculture, marine, pêche	0,0
Bâtiment, travaux publics	1,2
Électricité, électronique	0,1
Mécanique, travail des métaux	0,1
Industries de process	0,5
Matériaux souples, bois, industries graphiques	0,1
Maintenance	1,6
Ingénieurs et cadres de l'industrie	0,1
Transports, logistique et tourisme	1,0
Artisanat	0,1
Gestion, administration des entreprises	0,0
Informatique et télécommunications	0,3
Études et recherche	0,4
Administration publique, professions juridiques, armée et police	1,1
Banque et assurances	0,2
Commerce	1,4
Hôtellerie, restauration, alimentation	1,3
Services aux particuliers et aux collectivités	0,1
Communication, information, art et spectacle	0,5
Santé, action sociale, culturelle et sportive	1,1
Enseignement, formation	0,1
Dissimilarité de Duncan	**11,3**

Note : l'indicateur de Duncan est calculé sur l'ensemble des domaines professionnels (FAP 22) en dehors de celui de la politique et de la religion.

Lecture : l'indicateur de dissimilarité de Duncan indique que 11,3 % des offres STMT obtenues avec la matrice de passage secteurs-métiers (dont 1,1 % dans la santé, l'action sociale, culturelle et sportive) devraient changer de domaine professionnel pour que leur répartition par domaine s'aligne sur celle observée Le domaine du bâtiment est sous-représenté (en rouge), alors que celui de la santé, de l'action sociale, culturelle et sportive est surreprésenté par le recours à la matrice de passage (en jaune).

Champs : France (hors Mayotte), 2019.

Source : STMT (Dares, Pôle emploi), calculs Dares.

ANNEXE 2 - Dissimilarité par domaine professionnel entre la distribution des offres de la base Jocas et celle des recrutements avec diffusion d'une annonce en ligne

L'enquête Offre d'emploi et recrutement (Ofer) constitue une des sources comparables à la base Jocas. Elle porte sur les recrutements en CDI ou en CDD de plus d'un mois entre septembre et novembre 2015 des établissements d'au moins un salarié du secteur concurrentiel, à l'exception des domaines professionnels suivants : agriculture, marine, pêche ; administration publique, professions juridiques, armée et police ; politique, religion. Le champ des métiers est donc plus restreint que pour les autres sources ayant été comparées à la base Jocas. De plus, l'enquête date de 2015, ce qui fait un décalage de quatre années par rapport à la base Jocas et aux autres sources de données exploitées dans ce document. Cependant, Ofer permet d'identifier les recrutements ayant fait l'objet d'une diffusion d'une annonce en ligne : ce champ restreint « Ofer en ligne » est *a priori* le plus proche du champ Jocas. Pour ces raisons, la comparaison entre Jocas et Ofer est présentée de manière séparée par rapport à ce qui est fait entre Jocas et les autres sources.

Dans l'ensemble, les résultats sont similaires à ceux obtenus avec les autres sources (Tableau A2, Tableau 6). Les métiers systématiquement surreprésentés par Jocas par rapport aux autres sources le sont également ici, à l'exception du domaine du commerce. C'est aussi le cas pour les métiers sous-représentés par Jocas, sauf dans deux métiers de l'industrie (l'électricité, électronique et les industries de process). Ces différences pourraient s'expliquer par les différences de champs et de concepts mesurés ou par le décalage dans le temps entre les deux sources, étant données les évolutions rapides de la diffusion d'offres d'emploi en ligne.

TABLEAU A2 – Dissimilarité entre la base Jocas et les recrutements en ligne de l'enquête Ofer

En %

Domaine professionnel (FAP 22)	Recrutements avec diffusion d'une annonce en ligne
Bâtiment, travaux publics	3,1
Maintenance	1,3
Ingénieurs et cadres de l'industrie	0,7
Gestion, administration des entreprises	0,6
Informatique et télécommunications	2,2
Études et recherche	0,5
Commerce	2,0
Agriculture, marine, pêche	
Électricité, électronique	0,3
Industries de process	0,8
Matériaux souples, bois, industries graphiques	0,1
Transports, logistique et tourisme	0,9
Artisanat	0,1
Hôtellerie, restauration, alimentation	2,2
Services aux particuliers et aux collectivités	2,9
Communication, information, art et spectacle	0,4
Santé, action sociale, culturelle et sportive	2,0
Mécanique, travail des métaux	1,3
Administration publique, professions juridiques, armée et police	
Banque et assurances	0,1
Enseignement, formation	0,1
Dissimilarité de Duncan	**21,5**

Note : l'indicateur de Duncan est calculé sur l'ensemble des domaines professionnels (FAP 22) en dehors de ceux-ci : agriculture, marine, pêche ; administration publique, professions juridiques, armée et police ; politique et religion.

Lecture : l'indicateur de Duncan indique que 21,5 % des offres Jocas (dont 2,2 % dans l'informatique et les télécommunications) sont à redistribuer pour avoir une distribution similaire à celle des recrutements avec diffusion d'une annonce en ligne. Dans l'informatique et les télécommunications, la base Jocas compte relativement plus d'offres que l'enquête Ofer ne recense de recrutements avec offres diffusées en ligne (en jaune). Le domaine des services aux particuliers et aux collectivités est sous-représenté dans les offres diffusées sur le champ Jocas (en rouge).

Champs : Jocas : offres scrapées auprès d'une quinzaine de sites diffuseurs d'offres d'emploi en ligne en 2019. Ofer : ensemble des nouveaux recrutements en CDI ou en CDD de plus d'un mois entre septembre et novembre 2015 des établissements d'au moins un salarié du secteur concurrentiel à l'exception des domaines professionnels agriculture, marine, pêche, administration publique, professions juridiques, armée et police et politique, religion ; France.

Source : Dares, base Jocas 2019 et enquête Ofer 2016.

ANNEXE 3 – Dissimilarité de Duncan par famille professionnelle (Fap 87)

Le tableau A3 reproduit le tableau 6, représentant la décomposition de l'indice de dissimilarité de Duncan par domaine professionnel, au niveau plus détaillé des familles professionnelles. Il permet d'identifier les familles professionnelles à l'intérieur des domaines qui contribuent le plus à la dissimilarité entre la base Jocas et les autres sources.

TABLEAU A3 - Dissimilarité de Duncan par famille professionnelle (Fap 87)

En %

Domaine professioinel (FAP 22)	Offres STMT et partenaires	Emplois de moins d'un an (EEC)	Projets de recrutement (BMO)	Emplois vacants (Acemo)*	Intentions d'embauches (DPAE)*	Embauches (MMO)*
Agriculteurs, éleveurs, sylviculteurs, bûcherons	0,03	0,31	1,66	0,00	0,06	0,01
Maraîchers, jardiniers, viticulteurs	0,37	0,50	3,30	0,05	0,54	0,10
Techniciens et cadres de l'agriculture	0,02	0,10	0,01	0,01	0,00	0,01
Marins, pêcheurs, aquaculteurs	0,01	0,06	0,10	0,02	0,02	0,01
Ouvriers non qualifiés du gros œuvre du bâtiment, des travaux publics, du béton et de l'extraction	0,08	0,25	0,08	0,26	0,39	0,11
Ouvriers qualifiés des travaux publics, du béton et de l'extraction	0,04	0,03	0,08	0,00	0,13	0,05
Ouvriers qualifiés du gros œuvre du bâtiment	0,03	0,09	0,06	0,20	0,13	0,20
Ouvriers non qualifiés du second œuvre du bâtiment	0,02	0,05	0,04	0,16	0,17	0,02
Ouvriers qualifiés du second œuvre du bâtiment	0,01	0,60	0,41	0,53	0,35	0,83
Conducteurs d'engins du bâtiment et des travaux publics	0,06	0,11	0,09	0,07	0,06	0,12
Techniciens et agents de maîtrise du bâtiment et des travaux publics	0,03	0,45	0,56	0,37	0,67	0,73
Cadres du bâtiment et des travaux publics	**1,15**	**1,18**	**1,36**	**1,04**	**1,43**	**1,36**
Ouvriers non qualifiés de l'électricité et de l'électronique	0,02	0,05	0,05	0,04	0,10	0,00
Ouvriers qualifiés de l'électricité et de l'électronique	0,04	0,06	0,03	0,05	0,02	0,08
Techniciens et agents de maîtrise de l'électricité et de l'électronique	0,00	0,05	0,07	0,12	0,01	0,06
Ouvriers non qualifiés travaillant par enlèvement ou formage de métal	0,03	0,03	0,00	0,02	0,12	0,07
Ouvriers qualifiés travaillant par enlèvement de métal	0,01	0,23	0,19	0,20	0,15	0,27
Ouvriers qualifiés travaillant par formage de métal	0,02	0,06	0,02	0,03	0,17	0,12
Ouvriers non qualifiés de la mécanique	0,04	0,12	0,17	0,09	0,50	0,05
Ouvriers qualifiés de la mécanique	0,04	0,07	0,10	0,01	0,14	0,11
Techniciens et agents de maîtrise des industries mécaniques	0,18	0,22	0,38	0,03	0,26	0,35
Ouvriers non qualifiés des industries de process	0,09	0,14	0,47	0,06	1,13	0,10
Ouvriers qualifiés des industries de process	0,03	0,31	0,10	0,23	0,48	0,19
Techniciens et agents de maîtrise des industries de process	0,05	0,19	0,36	0,19	0,28	0,33
Ouvriers non qualifiés du textile et du cuir	0,04	0,02	0,12	0,02	0,01	0,01
Ouvriers qualifiés du textile et du cuir	0,04	0,03	0,02	0,07	0,02	0,01
Ouvriers non qualifiés du travail du bois et de l'ameublement	0,01	0,12	0,03	0,12	0,10	0,04
Ouvriers qualifiés du travail du bois et de l'ameublement	0,01	0,09	0,02	0,12	0,11	0,04
Ouvriers des industries graphiques	0,06	0,01	0,01	0,02	0,06	0,03
Techniciens et agents de maîtrise des matériaux souples, du bois et des industries graphiques	0,03	0,02	0,03	0,00	0,02	0,02
Ouvriers qualifiés de la maintenance	0,00	0,18	0,03	0,15	0,09	0,23
Ouvriers qualifiés de la réparation automobile	0,10	0,37	0,34	0,39	0,51	0,45
Techniciens et agents de maîtrise de la maintenance	0,36	1,70	1,98	1,42	1,80	1,82
Ingénieurs et cadres techniques de l'industrie	0,71	1,00	1,26	0,91	1,32	1,26

Domaine professioinel (FAP 22)	Offres STMT et partenaires	Emplois de moins d'un an (EEC)	Projets de recrutement (BMO)	Emplois vacants (Acemo)*	Intentions d'embauches (DPAE)*	Embauches (MMO)*
Ouvriers non qualifiés de la manutention	0,27	0,72	0,48	0,35	2,40	0,54
Ouvriers qualifiés de la manutention	0,15	0,41	0,07	0,35	1,39	0,38
Conducteurs de véhicules	0,04	0,29	0,41	0,21	0,17	0,02
Agents d'exploitation des transports	0,04	0,40	0,42	0,39	0,41	0,44
Agents administratifs et commerciaux des transports et du tourisme	0,07	0,17	0,13	0,10	0,29	0,11
Cadres des transports, de la logistique et navigants de l'aviation	0,15	0,05	0,19	0,11	0,15	0,16
Artisans et ouvriers artisanaux	0,01	0,19	0,01	0,18	0,16	0,17
Secrétaires	0,13	0,46	0,56	0,39	0,51	0,40
Employés de la comptabilité	0,35	0,09	0,29	0,20	0,18	0,12
Employés administratifs d'entreprise	0,47	0,28	0,47	0,44	0,55	0,57
Secrétaires de direction	0,16	0,47	0,58	0,42	0,52	0,51
Techniciens des services administratifs, comptables et financiers	0,07	0,68	0,49	0,74	0,33	0,25
Cadres des services administratifs, comptables et financiers	**1,34**	**2,04**	**2,66**	**1,86**	**2,41**	**2,31**
Dirigeants d'entreprises	0,13	0,23	0,22	0,23	0,23	0,23
Employés et opérateurs de l'informatique	0,09	0,03	0,07	0,12	0,05	0,05
Techniciens de l'informatique	0,22	0,15	0,01	0,40	0,17	0,12
Ingénieurs de l'informatique	**1,73**	**2,00**	**2,13**	**1,47**	**2,68**	**2,35**
Personnels d'études et de recherche	0,51	0,27	0,46	0,14	0,65	0,59
Employés administratifs de la fonction publique (catégorie C et assimilés)	0,00	0,65	0,01	0,47	0,58	0,71
Professions intermédiaires administratives de la fonction publique (catégorie B et assimilés)	0,00	0,28	0,00	0,13	0,21	0,28
Cadres de la fonction publique (catégorie A et assimilés)	0,11	0,00	0,23	0,06	0,06	0,03
Professionnels du droit (hors juristes en entreprise)	0,00	0,04	0,04	0,01	0,06	0,05
Armée, police, pompiers	0,05	0,12	0,03	0,02	0,04	0,01
Employés de la banque et des assurances	0,11	0,26	0,08	0,20	0,02	0,07
Techniciens de la banque et des assurances	0,00	0,03	0,06	0,00	0,19	0,17
Cadres de la banque et des assurances	0,09	0,14	0,10	0,16	0,08	0,06
Caissiers, employés de libre service	0,53	0,40	1,42	0,38	0,07	0,24
Vendeurs	0,20	1,78	0,44	1,51	0,54	0,90
Attachés commerciaux et représentants	0,22	0,29	0,84	0,09	0,76	0,52
Maîtrise des magasins et intermédiaires du commerce	0,07	0,19	0,21	0,29	0,09	0,18
Cadres commerciaux et technico-commerciaux	**0,97**	**2,76**	**2,98**	**2,47**	**3,12**	**2,95**
Bouchers, charcutiers, boulangers	0,11	0,16	0,06	0,18	0,11	0,00
Cuisiniers	0,22	0,21	1,32	0,30	0,14	0,78
Employés et agents de maîtrise de l'hôtellerie et de la restauration	**0,37**	**1,04**	**1,64**	**1,81**	**1,53**	**2,69**
Patrons et cadres d'hôtels, cafés, restaurants	0,29	0,03	0,15	0,01	0,04	0,04
Coiffeurs, esthéticiens	0,00	0,04	0,01	0,21	0,12	0,03
Employés de maison	0,39	0,28	0,05	0,45	0,39	0,44
Aides à domicile et aides ménagères	0,31	0,79	0,91	0,85	0,60	1,47
Assistantes maternelles	1,12	0,07	0,45	0,18	0,32	0,10
Agents de gardiennage et de sécurité	0,08	0,29	0,43	0,01	1,01	0,16
Agents d'entretien	**0,20**	**1,48**	**2,05**	**0,75**	**3,05**	**1,83**
Employés des services divers	0,01	0,07	0,16	0,07	0,11	0,20

Domaine professioinel (FAP 22)	Offres STMT et partenaires	Emplois de moins d'un an (EEC)	Projets de recrutement (BMO)	Emplois vacants (Acemo)*	Intentions d'embauches (DPAE)*	Embauches (MMO)*
Professionnels de la communication et de l'information	0,08	0,21	0,14	0,21	0,24	0,74
Professionnels des arts et des spectacles	0,14	0,45	1,72	0,58	1,15	2,54
Aides-soignants	0,03	0,79	0,60	0,37	0,35	1,15
Infirmiers, sages-femmes	0,03	0,01	0,04	0,13	0,06	0,37
Médecins et assimilés	0,03	0,01	0,08	0,02	0,06	0,19
Professions para-médicales	0,01	0,07	0,01	0,02	0,11	0,12
Professionnels de l'action sociale et de l'orientation	0,09	0,31	0,13	0,23	0,11	0,60
Professionnels de l'action culturelle, sportive et surveillants	0,22	0,81	2,07	0,53	0,88	1,95
Enseignants	0,60	0,22	0,14	0,11	0,15	0,02
Formateurs	0,02	0,01	0,04	0,09	0,03	0,19
Dissimilarité de Duncan	**16,4**	**32,1**	**41,8**	**28,0**	**41,0**	**40,3**

Note : l'indicateur de Duncan est calculé sur l'ensemble des familles professionnelles (Fap 87) en dehors de celui de la politique et du clergé.

* Sources non déclinables par métier et pour lesquelles une matrice de passage secteur – métier a été utilisée pour passer du secteur NAF 38 aux familles professionnelles FAP 87.

Lecture : 41,8 % des offres scrapées en ligne (dont 2,13 % chez les ingénieurs de l'informatique) devraient changer de famille professionnelle pour que la répartition de leurs familles professionnelles s'aligne sur celle des projets de recrutement issus de l'enquête BMO. La plus importante contribution à cette dissimilarité est due à la famille « Maraîchers, jardiniers, viticulteurs » (3,30 points sur 41,8 %). Les cadres du bâtiment et des travaux publics comptent relativement plus d'offres scrapées que de projets de recrutement (en jaune), alors que les agents d'entretien comptent relativement plus de projets de recrutement (en rouge).

Champs : France (hors Mayotte), 2019.

Source : Jocas, STMT, offres partenaires, BMO (Pôle emploi), EEC (Insee), DPAE (Urssaf) et Acemo (Dares) - calculs Dares.