Journées de méthodologie statistique de l'Insee 2022

ENJEUX STATISTIQUES DU SYSTÈME D'INFORMATION INSERJEUNES SUR L'INSERTION PROFESSIONNELLE

Nathalie CARON (*), Loïc MIDY (**)

(*) DEPP, Ministère de l'Éducation nationale
(**) DEPP[1], Ministère de l'Éducation nationale

nathalie.caron@education.gouv.fr
loic.midy@insee.fr

Mots-clés *(6 maximum)* : appariement, sources administratives, modèles multiniveaux, insertion professionnelle des jeunes.

Domaine concerné : Appariement et combinaison de sources, données administratives, appariement.

Résumé

La Direction de l'Evaluation, de la Prospective et de la Performance (Depp) réalise depuis le début des années 90 deux enquêtes d'insertion annuelles permettant de suivre l'entrée dans la vie active des jeunes sortant d'apprentissage ou de voie professionnelle scolaire. Mais celles-ci ne permettent pas de publier des statistiques au niveau établissement, centre de formation d'apprentis et lycée professionnel, comme requis par la loi pour la liberté de choisir son avenir professionnel de 2018. Afin de répondre à ce besoin, la Depp et la Direction de l'Animation de la Recherche, des Etudes et des Statistiques (Dares) ont construit un nouveau dispositif, InserJeunes, basé sur l'appariement d'une vingtaine de sources administratives principalement sur identifiants indirects. En particulier, les appariements des bases des jeunes sortants du système éducatif avec la source Mouvement de Main d'Œuvre (MMO) de la Dares basée sur la Déclaration Sociale Nominative (DSN) permettent de savoir s'ils ont un emploi salarié en France. Le premier enjeu statistique de ce dispositif consiste à définir une méthodologie d'appariements robuste et rapide qui ne nécessite pas de reprise manuelle par des gestionnaires. Dans InserJeunes, chaque processus d'appariement comporte cinq étapes (normalisation des données, indexation, calcul de similarités, classification supervisée et évaluation de la qualité) et est réalisé via un outil d'appariement spécifique développé par la Depp. Le second enjeu statistique consiste à accompagner la publication des taux d'emploi par établissement par un indicateur de « valeur ajoutée » permettant de mesurer l'apport propre de l'établissement. En effet, une simple comparaison des taux d'emploi calculés entre établissements ne suffit pas car l'insertion dépend en partie de facteurs extérieurs à l'établissement : profil des jeunes, type et niveau de formation, spécialité de formation et marché du travail local. La notion de valeur ajoutée est déjà utilisée par la Depp lors de la diffusion chaque année des taux de réussite au baccalauréat par lycée (indicateurs de valeur ajoutée des lycées – IVAL-). Étant donné la mobilisation de nombreuses données à caractère personnel (ex : noms, prénoms, identifiant national propre à chaque élève, étudiant ou apprenti qui est spécifique à la sphère éducation – INE-), l'aspect « sécurité » ainsi que la traçabilité complète des traitements réalisés sont également des enjeux importants pour le dispositif InserJeunes.

[1]Depp lors de la rédaction de l'article

Abstract

Since the beginning of the nineties, the direction of evaluation, prospective and performance (Depp) carries out two annual surveys on the labour market integration of the students who just finished their study as apprentice or in vocational school path. But they don't enable to publish statistics at the establishment level as required by the 2018 law for the freedom to choose one professional future. So, the Depp and the direction of animation, research, studies and statistics (Dares) have designed a new information system, InserJeunes, based on the record linkage of administrative data sources. he first statistical challenge of this information system is to define a robust and quick record linkage methodology. A specific tool was developed by the Depp. The second statistical challenge is to calculate an "added value" indicator to measure the establishment's own contribution. The last challenge is to protect the data against fraudulent access since there are many personal data in the system such as first and last names and to have complete traceability of the processing carried out.

---

1. Introduction

L'orientation des élèves se construit tout au long de la scolarité avec en particulier des étapes clés en fin de classes de troisième, seconde générale et technologique et terminale. Ainsi, l'orientation en voie professionnelle peut commencer dès la fin de troisième avec un choix entre apprentissage ou voie professionnelle scolaire. L'insertion dans l'emploi étant la première finalité de la formation professionnelle, connaître les taux d'insertion des formations initiales permet d'éclairer les choix des jeunes et de leur famille.

La Direction de l'Evaluation, de la Prospective et de la Performance (Depp) a mis en place au début des années 1990, deux enquêtes d'insertion exhaustives annuelles[2] réalisées directement auprès des jeunes permettant de suivre l'entrée dans la vie active des sortants d'apprentissage et de voie professionnelle scolaire 7 mois après leur sortie du système éducatif. Ces dernières apportent des informations précieuses mais ne permettent pas de diffuser des statistiques à un niveau fin que ce soit en termes de niveau géographique ou en termes de précision du diplôme préparé compte tenu des taux de réponse de l'ordre de 60%. La diffusion principale des résultats portait sur le niveau national par diplôme et regroupement de spécialités et certaines académies faisaient également quelques diffusions de résultats. Or, la loi du 5 septembre 2018 pour la liberté de choisir son avenir professionnel prévoit en son article 24 la publication de statistiques sur le parcours scolaire[3] (taux d'obtention des diplômes ou titres professionnels, taux de poursuite d'études, taux d'interruption en cours de formation) et le taux d'insertion dans l'emploi des jeunes en formation professionnelle pour chaque centre de formation d'apprentis et pour chaque lycée professionnel. La loi précise que le taux d'emploi par établissement est accompagné d'une « valeur ajoutée » qui permet de comparer de façon pertinente le taux d'emploi des élèves sortants d'un établissement au taux d'emploi d'établissements similaires et de mesurer l'effet propre de l'établissement concerné. Le concept de « valeur ajoutée » existe déjà dans la diffusion des taux d'obtention du baccalauréat par lycée publiée chaque année par la Depp (indicateurs IVAL - voir [5]).

Afin de répondre à ce besoin nouveau auquel ne pouvaient répondre les enquêtes d'insertion annuelles, la Depp et la Direction de l'Animation de la Recherche, des Etudes et des Statistiques

[2] Les enquêtes « Insertion dans la vie active (IVA) » et « Insertion professionnelle des apprentis (IPA) »

[3] Article 24 loi n° 2018-771 : « chaque année, pour chaque centre de formation d'apprentis et pour chaque lycée professionnel, sont rendus publics quand les effectifs concernés sont suffisants :
1° Le taux d'obtention des diplômes ou titres professionnels 2° Le taux de poursuite d'études 3° Le taux d'interruption en cours de formation 4° Le taux d'insertion professionnelle des sortants de l'établissement concerné, à la suite des formations dispensées 5° La valeur ajoutée de l'établissement. Pour chaque centre de formation d'apprentis, est également rendu public chaque année le taux de rupture des contrats d'apprentissage conclus. Les modalités de diffusion des informations publiées sont déterminées par arrêté conjoint des ministres chargés de la formation professionnelle et de l'éducation nationale ».

(Dares – ministère du travail) ont construit un nouveau dispositif, appelé InserJeunes, basé sur l'appariement de sources administratives exhaustives relatives à la scolarité des élèves et apprentis, à la réussite aux examens, aux contrats d'apprentissage et aux contrats salariés (dispositif mouvement de main d'œuvre – MMO - basée sur la déclaration sociale nominative – DSN -). InserJeunes couvre deux champs sur l'ensemble de la France hors Mayotte[4] : les apprentis préparant une certification de niveau 3[5] (ex : CAP), 4 (ex : Brevet Professionnel) ou 5 (ex : BTS) et les élèves de voie professionnelle scolaire relevant du ministère chargé de l'éducation nationale des secteurs public et privé sous contrat. Le champ des élèves de voie professionnelle scolaire relevant du ministère chargé de l'Agriculture est également intégré dans InserJeunes mais pour l'instant à titre expérimental, ces données ne donnant pas encore lieu à des statistiques officielles[6].

Pour chaque centre de formation d'apprentis, est également rendu public chaque année le taux de rupture des contrats d'apprentissage conclus, indicateur indiqué dans la loi du 5 septembre 2018 également et calculé par la Dares en dehors d'InserJeunes.

L'ensemble de ces indicateurs qui sont complémentaires et disponibles à un niveau fin constituent des éléments d'information sur l'entrée dans la vie active pour éclairer les choix des jeunes et de leur famille sur l'enseignement professionnel ainsi que des outils de pilotage mis à disposition des chefs d'établissement. De façon plus générale, ils peuvent également contribuer aux réflexions dans chaque territoire sur la formation professionnelle et l'emploi. Avec des informations détaillées sur les contrats salariés (type de contrat, salaire, quotité de travail, catégorie socio-professionnelle...) et sur l'établissement employeur (secteur, commune d'implantation, …) issues de la DSN ainsi que sur la formation des jeunes issue des sources exhaustives relatives à la scolarité, InserJeunes est également une source très riche qui permet de réaliser de nombreuses études notamment pour éclairer les thématiques de l'adéquation formation / emploi, les conditions d'emploi de jeunes sortants de formation et les disparités territoriales sur l'insertion professionnelle des jeunes (voir [2], [3] et [4]).

Pour développer InserJeunes, plusieurs défis méthodologiques ont dû être relevés. Etant donné le nombre de bases à apparier sur identifiant indirect, le premier d'entre eux a été de définir une méthodologie d'appariements ne nécessitant pas de reprise manuelle par des gestionnaires ainsi qu'un outil d'appariement adapté (c'est-à-dire notamment rapide et générique). Le second a consisté à calculer des indicateurs de « valeur ajoutée » pour chaque établissement (centre de formation d'apprentis - CFA - et lycée professionnel) permettant de mesurer l'apport propre de l'établissement en s'inspirant de la méthodologie développée dans le cadre des IVAL tout en la complétant car les facteurs extérieurs à l'établissement peuvent être liés à d'autres facteurs que ceux en lien avec les formations suivies par les jeunes comme le marché du travail local. Enfin, étant donné la mobilisation de nombreuses données à caractère personnel et les flux de données entre la Depp et la Dares, l'aspect « sécurité » et le respect du RGPD ont également été des éléments cruciaux du dispositif InserJeunes. Par conséquent, la traçabilité complète des traitements réalisés est importante, ce qui a été facilité par le fait que l'ensemble des traitements sont pilotés directement par l'équipe de production statistique via une application idoine.

Dans la première partie de cet article, nous développerons les choix opérés dans le cadre du projet ainsi que les grands principes du processus InserJeunes et ses limites. La seconde partie abordera l'élément technique essentiel qu'est le moteur d'appariement utilisé. La troisième partie détaillera l'indicateur portant sur le taux d'emploi ainsi que la « valeur ajoutée » permettant d'estimer l'apport propre de chaque établissement. Enfin, dans la dernière partie, seront abordées les enjeux de sécurité, les bases de données traitées contenant des données à caractère personnel ainsi que les bonnes pratiques et innovations mises en œuvre dans InserJeunes.

[4] Les emplois salariés à Mayotte n'ont pas encore complètement basculé en DSN (déclaration sociale nominative).

[5] La nomenclature des diplômes par niveau utilisée dans cet article est celle du décret n° 2019-14 du 8 janvier 2019 relatif au cadre national des certifications professionnelles.

[6] Il reste des travaux de mise en qualité statistique à réaliser avant de pouvoir produire des statistiques officielles sur ce champ.

2. Les grands principes du processus statistique InserJeunes et ses limites

2.1. Les choix opérés

Le nouveau dispositif sous co-maitrise d'ouvrage Depp-Dares, appelé InserJeunes, basé sur l'appariement de sources administratives exhaustives relatives à la scolarité des élèves et apprentis, à la réussite aux examens, aux contrats d'apprentissage et aux contrats salariés de la déclaration sociale nominative (la DSN) a été développé pour permettre de répondre à la loi du 5 septembre 2018 pour la liberté de choisir son avenir professionnel qui prévoyait en son article 24 la publication de statistiques sur le parcours scolaire (taux de poursuite d'études, taux d'interruption en cours de formation) et le taux d'insertion dans l'emploi des jeunes en formation professionnelle pour chaque centre de formation d'apprentis et pour chaque lycée professionnel. Ce dernier indicateur donne la part, parmi les élèves ne poursuivant pas leurs études, de ceux qui disposent d'un contrat de travail enregistré dans la DSN, sur une semaine de référence, 6 mois (et puis 12, 18 ou 24 mois) suivant la fin de leurs études. Le mois de référence retenu porte alternativement sur janvier (mesure de l'emploi à 6 et 18 mois) et sur juillet (mesure de l'emploi à 12 et 24 mois). Plus précisément, les semaines de référence retenues sont la deuxième de janvier et la première de juillet de façon à éviter respectivement le 1er janvier et le 14 juillet. Pour mémoire, la DSN est obligatoire pour tout employeur du secteur privé du régime général et du régime agricole de Sécurité sociale depuis le 1er janvier 2017, elle constitue de fait une source quasi exhaustive pour l'emploi salarié dans le secteur privé.

L'indicateur concernant « le taux d'obtention des diplômes » évoqué également dans l'article 24 de la loi n'est pas calculé dans InserJeunes mais est disponible dans le cadre des indicateurs de valeur ajoutée des lycées (IVAL) diffusés par la Depp depuis 1993 (voir [5] et [6]). Les IVAL sont actuellement calculés pour l'ensemble des lycées d'enseignement général et technologique et des lycées professionnels, publics et privés sous contrat et portent uniquement sur l'obtention du baccalauréat. Ces indicateurs ont vocation à s'étendre aux autres diplômes ainsi qu'à d'autres champs d'établissement. Dans le cadre des IVAL, trois indicateurs sont calculés pour mesurer la capacité des lycées à accompagner leurs élèves jusqu'au baccalauréat, le taux de réussite, le taux d'accès qui évalue la probabilité pour un élève d'obtenir le baccalauréat à l'issue d'une scolarité entièrement effectuée dans le lycée, même s'il y a redoublé et le taux de mentions au baccalauréat. Au-delà du seul taux de réussite à l'examen, les « valeurs ajoutées » associées aux indicateurs bruts facilitent les comparaisons entre des établissements hétérogènes, en prenant en compte les disparités scolaires et socio-économiques entre lycées (voir [5]). L'indicateur « taux de rupture des contrats d'apprentissage conclus » également évoqué dans l'article 24 pour les centres de formations d'apprentis est quant à lui calculé par la Dares et diffusé par InserJeunes. La suite de cet article n'abordera donc pas ces deux indicateurs.

InserJeunes permet de se rapprocher de l'exhaustivité et de construire des indicateurs à des niveaux très fins. Ainsi, même si la loi stipulait que le calcul des indicateurs devrait se faire au niveau établissement, il a été décidé dès le début du projet d'avoir ces indicateurs déclinés par diplôme et regroupement de spécialités au sein de chaque établissement, voire diplôme fin si l'effectif était suffisant. En effet, le taux d'emploi augmentant avec le niveau de diplôme, un même taux d'emploi au niveau établissement s'apprécie en fonction du niveau et des spécialités des formations offertes par l'établissement.

Afin de pouvoir diffuser le plus de croisements possibles au niveau établissement, niveau et diplôme préparé, le calcul de indicateurs et la diffusion au niveau des établissements ont été réalisés sur un cumul de deux années de jeunes sortants de formation. De plus, pour atténuer les variations qui peuvent être fortes d'une année sur l'autre, pour des établissements aux petits effectifs notamment, et ainsi maintenir un certain niveau de fiabilité tout en préservant le secret statistique, les taux calculés ne sont pas renseignés lorsque les effectifs sont trop faibles pour être représentatifs (moins

de 20 au dénominateur sur les deux années cumulées). La première diffusion des résultats issus de ce nouveau système d'information qui ont été diffusés en février 2021 portait au niveau établissement sur les deux millésimes cumulés suivants : les jeunes inscrits en 2017-2018 et 2018-2019 pour le taux de poursuite d'études et les sortants d'études en 2018 et 2019 pour les autres indicateurs.

Le développement de ce nouveau dispositif a duré presque trois années et les premiers résultats ont été diffusés en février 2021. Il a pu bénéficier d'un financement du Fonds pour la Transformation de l'Action Publique (FTAP), ce qui a permis d'avoir une équipe dédiée alliant compétences statistiques et informatiques qui a ainsi développé en interne tous les programmes et qui a bénéficié des compétences métiers sur les différentes sources utilisées présentes dans les autres structures de la Depp et de la Dares.

2.2. Le processus InserJeunes

Le processus principal d'InserJeunes contient plusieurs étapes de traitement articulées sur une vingtaine de bases nécessaires au calcul des différents indicateurs : bases sur les élèves de la Depp, de la sous-direction des Systèmes d'Information et des Etudes Statistiques (SIES, ministère en charge de l'enseignement supérieur) ainsi que celles des élèves des établissements agricoles de la Direction Générale de l'Enseignement et de la Recherche (DGER, ministère en charge de l'agriculture), bases sur les résultats aux examens de la Depp, du SIES et de la DGER ainsi que la base Mouvement de Main d'œuvre basée sur la DSN de la Dares appelée « DSN » dans la suite de cet article (voir figure 1). Chaque rapprochement entre ces bases se fait soit sur l'identifiant spécifique à la sphère éducation « l'identifiant national élève (INE[7]) » qui est une donnée à caractère personnel indirectement identifiante, lorsque celui-ci est présent dans les bases « élèves », soit sur des variables directement identifiantes (nom, prénom, sexe, date et lieu de naissance, …), soit dans certains cas sur un mixte de ces deux méthodes lorsque l'INE est partiellement rempli.

L'intensité des méthodes selon les différentes étapes de traitement est représentée par des flèches plus ou moins larges dans la figure 1.

Dans l'étape 1, le champ des élèves en année terminale de formation une année scolaire donnée est calculé en mobilisant trois bases de données administratives « scolarité »[8] chacune couvrant une partie du champ InserJeunes : l'apprentissage, la voie professionnelle scolaire dans un établissement du ministère en charge de l'éducation nationale et celle dans un établissement du ministère en charge de l'agriculture. Ces bases contiennent des variables directement identifiantes (nom, prénom, date de naissance, …), l'identifiant national élève (INE) ainsi que des informations sur l'établissement et la formation suivie.

Dans l'étape 2, le champ des élèves sortants c'est-à-dire ceux qui ne sont plus en formation dans le système éducatif est établi en recherchant, principalement sur l'INE, si ces élèves sont présents l'année scolaire suivante dans l'ensemble des bases de données élèves disponibles c'est-à-dire dans les trois mêmes bases ainsi que dans trois bases supplémentaires[9] permettant d'être le plus exhaustif possible[10] sur les poursuites d'études. Tout élève retrouvé, qu'il redouble, qu'il poursuive ses études ou s'oriente vers une autre formation de tout niveau, est noté comme étant toujours en études, les

[7] Mis en place à partir de 2017, cet identifiant national (INE) propre à chaque élève, étudiant ou apprenti a vocation à faciliter la gestion du système éducatif et à permettre le suivi statistique des élèves, ,étudiants et des apprentis.

[8] SIFA (système d'information de la formation des apprentis) pour les apprentis, SYSCA-SCO (Système d'information Statistique Consolidé Académique) pour les élèves de voie professionnelle scolaire du ministère en charge de l'éducation nationale et DeciEA pour les élèves de voie professionnelle scolaire du ministère en charge de l'agriculture.

[9] En plus des bases au cœur du processus (SIFA, SYSCA-SCO, DeciEA), on prend également en compte les élèves du secteur privé hors contrat avec la source SCOLEGE et les étudiants dans l'enseignement supérieur avec les enquêtes SISE (Système d'information sur le suivi des étudiants) et les vœux validés dans ParcourSup dans un institut de formation en soins infirmiers.

[10] En particulier en prenant en compte les poursuites dans l'enseignement supérieur.

autres sont appelés les « sortants de formation[11] ». Cette information permet de calculer le taux de poursuite d'études qui mesure la part d'élèves toujours en formation en France l'année scolaire suivant leur dernière année dans les cursus suivis dans InserJeunes.

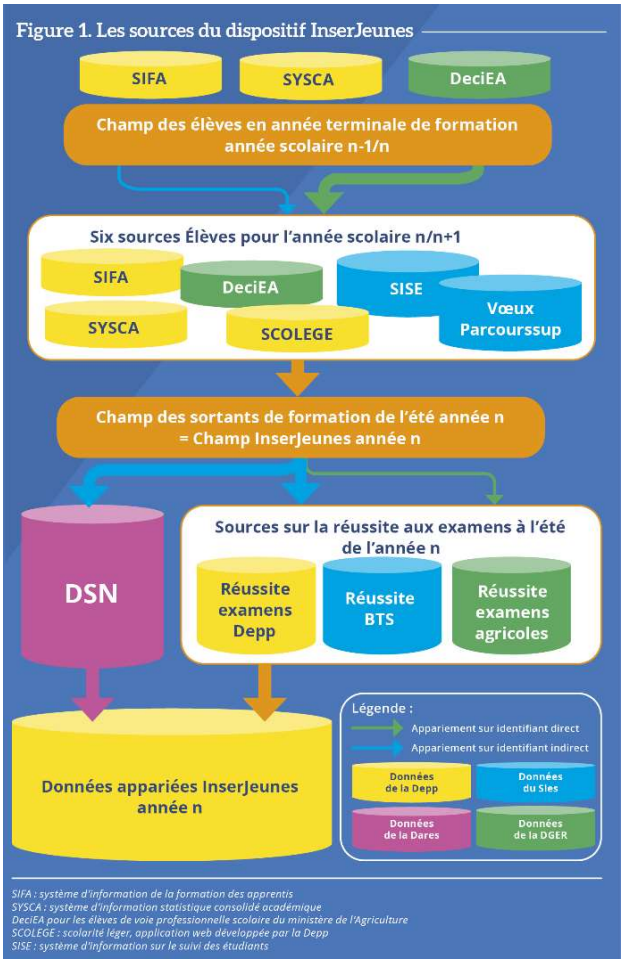
Dans l'étape 3, les bases élèves/apprentis sont enrichies avec leur réussite aux examens (selon les examens cet appariement est réalisé sur l'INE ou sur identifiants indirects) *ce qui permet de calculer le taux d'interruption en cours de formation.* Cet indicateur donne une estimation du risque d'interrompre sa formation sur l'ensemble de la durée du diplôme et s*on calcul repose sur une méthode conjoncturelle similaire à celle utilisée pour déterminer le taux d'accès au baccalauréat qui est un des indicateurs de valeur ajoutée (IVAL) établis par la Depp de longue date (voir [8])*. *Il est disponible* uniquement par diplôme et par établissement et seulement pour les diplômes pour lesquels nous disposons des bases de réussite aux examens des élèves[12].

Puis dans l'étape 4, les bases élèves/apprentis sortants sont appariées sur identifiant indirect (nom, prénom, date de naissance, …) *avec la DSN*[13] *ce qui permet de mesurer un taux d'emploi salarié en France des sortants puis* la valeur ajoutée de l'établissement sur ce taux d'emploi. Cette dernière source contient des informations détaillées sur les contrats salariés (type de contrat, salaire, quotité de travail, catégorie socioprofessionnelle...) ainsi que sur l'établissement employeur (secteur, commune d'implantation, …) ce qui permet d'utiliser également InserJeunes pour réaliser des études statistiques, par exemple, sur l'adéquation formation/emploi.

Les données permettant de calculer le taux de poursuite d'études ainsi que le taux d'emploi sont ensuite agrégées notamment au niveau établissement. Avant la publication des résultats, les chefs d'établissements (CFA et lycée professionnel) peuvent consulter les données de leur établissement via un site web dédié et remonter leurs observations dans un délai de 15 jours. Les remontées sont ensuite instruites et peuvent conduire à des ajustements des indicateurs.

Enfin dans la dernière étape, les données sont diffusées publiquement. La diffusion d'InserJeunes prend plusieurs formes : publication de *Notes d'Information*, diffusion de fichiers Excel sur le site du ministère de l'éducation nationale, mise en open data sur le site du ministère de l'éducation nationale[14], diffusion de données via le site web de diffusion grand public[15] et accès aux données via le web service de diffusion ouvert uniquement à des partenaires institutionnels[16].

[11] En réalité, quelques sortants de formation peuvent en fait être encore en études car nous n'avons pas les poursuites d'études à l'étranger.
[12] CAP, Mention complémentaires de niveau 3, Mentions complémentaires de niveau 4, Brevet professionnel (BP) pour les apprentis, baccalauréat professionnel et BTS.
[13] Pour être tout à fait précis, on utilise la source MMO (mouvement de main d'œuvre) basée sur la DSN.
[14] https://data.education.gouv.fr/pages/accueil
[15] https://www.inserjeunes.education.gouv.fr/diffusion/accueil
[16] Par exemple, l'Onisep et la DGESCO.



Cette figure est extraite de l'article de Midy, L. 2021, « un outil d'appariement sur identifiants indirects : L'exemple du système d'information sur l'insertion des jeunes », Courrier des statistiques N6 – 2021, Insee

2.3. Les limites d'InserJeunes

InserJeunes présente plusieurs limites.

Les poursuites d'études en dehors des sources répertoriées dans InserJeunes ne sont pas comptabilisées (notamment celles à l'étranger). Cela conduit à surestimer l'effectif de sortants de formations de manière très limitée[17] sauf pour les élèves de voie professionnelle scolaire dépendant du ministère de l'agriculture[18].

Etant donné qu'InserJeunes mesure l'emploi en se fondant sur la DSN, le champ de l'emploi couvert par InserJeunes est donc directement celui de la DSN. Or, il manque à ce jour en DSN les emplois non salariés, dans la fonction publique[19], chez les particuliers employeurs, une partie de l'emploi salarié agricole dans les entreprises de moins de 10 salariés et les emplois à l'étranger. Cela conduit à minorer l'emploi des sortants d'apprentissage d'environ 4 points de pourcentage en moyenne et l'emploi des sortants de voie professionnelle scolaire relevant du ministère chargé de l'éducation nationale des secteurs public et privé sous contrat d'environ 2 points de pourcentage. Ces estimations de sous-couverture ont été réalisées à partir des résultats des enquêtes IVA et IPA. Le défaut de couverture est plus important pour les sortants de voie professionnelle scolaire dépendant du ministère de l'agriculture ce qui explique (avec le fait que les sortants de cette voie sont surestimés) que ce champ reste en expérimentation dans InserJeunes pour le moment.

[17] Cette information est obtenue par comparaison entre InserJeunes et les enquêtes IVA et IPA
[18] La surestimation est d'environ 15% pour ce sous champ.
[19] Ils seront intégrés en 2023.

Enfin, la DSN n'étant pas pleinement déployée à Mayotte, les établissements de cette île ne font pas encore partie du champ d'InserJeunes.

3. Le moteur d'appariement : un élément essentiel

Le processus statistique InserJeunes comporte au total dix appariements sur identifiants indirects sur nom(s), prénom(s), sexe, date de naissance et lieu de naissance (si celui-ci est disponible dans la source) pour chaque année scolaire. La problématique des appariements sur identifiants indirects est donc un enjeu central de ce nouveau SI (voir [10]).

Il s'agit en fait d'appariements automatiques à finalité statistique, à distinguer d'appariements à finalité administrative. Dans les deux cas, on cherche à minimiser les erreurs, mais dans un appariement « statistique » on souhaite un équilibre entre les faux positifs (les personnes appariées à tort) et les faux négatifs (les personnes non appariées à tort) alors que dans un appariement « administratif » on souhaite en général minimiser une de ces deux catégories. Par exemple, si on souhaite radier les personnes décédées dans la base des permis de conduire on veut minimiser les faux positifs c'est-à-dire les personnes appariées à tort entre les deux bases soit les personnes encore vivantes mais qui sont radiées dans la base des permis de conduire du fait d'une erreur d'appariement. De plus, en général dans le cadre d'un appariement administratif, l'enjeu pouvant être important, des reprises de cas douteux sont souvent réalisées par des gestionnaires. C'est par exemple le cas lors de l'attribution ou la vérification de l'identifiant national (INE) propre à chaque élève, étudiant ou apprenti dans la sphère éducation où des gestionnaires dans les académies sont mobilisés tout au long de l'année afin d'assurer une qualité maximale à cet identifiant.

Enfin, comme le nombre d'appariements est relativement important, il est nécessaire de mettre au point un processus d'appariement général puis d'en faire une implémentation informatique générique et rapide.

3.1. Un appariement spécifique « qualité » pour contrôler le moteur

Dans les appariements réalisés dans InserJeunes, le taux d'appariement ne donne aucune indication du niveau de qualité du processus. Par exemple, lorsqu'un sortant n'est pas apparié avec la DSN, il n'est pas possible de savoir si c'est parce qu'il n'est pas en emploi salarié dans le secteur privé ou en raison d'une erreur dans le processus d'appariement. Mais, InserJeunes comporte un appariement sur identifiants indirects dit « qualité » annuel pour lequel le taux d'appariement théorique est de 100% : il s'agit de l'appariement du fichier recensant les apprentis au 31/12 ayant un contrat d'apprentissage actif[20] avec la DSN puisque les contrats d'apprentissage sont dans cette source. Ainsi, le taux d'appariement réel obtenu constitue un indicateur de la qualité du processus d'appariement.

3.2. Un processus d'appariement en cinq étapes

Dans InserJeunes, chaque appariement sur identifiants indirects est réalisé sur deux tables individuelles[21] sans double compte. Le processus d'appariement retenu pour InserJeunes comporte 5 étapes successives, comme dans la présentation de Peter Christen (voir [1] et figure 2).

Tout d'abord, les données sont normalisées. Les identifiants indirects utilisés dans l'appariement se présentent selon des formats hétérogènes dans les différentes sources mobilisées dans InserJeunes. La normalisation des données consiste à les recoder selon une structure commune (exemple : mettre les lettres en majuscule) afin de faciliter les traitements ultérieurs.

[20] En dehors de la fonction publique car les emplois publics ne sont pas encore intégrés en DSN.
[21] C'est-à-dire que l'unité d'observation est l'individu (ici élève ou apprenti).

Puis vient l'étape d'indexation des données qui consiste à établir une liste de taille raisonnable de paires « potentiellement intéressantes ». Une paire correspond au croisement d'une ligne de la première table avec une ligne de la seconde table. Chaque paire comporte donc un/des noms, un/des prénoms, une date de naissance, un lieu de naissance et une variable sexe provenant de chacune des deux tables qu'on apparie. Dans un premier temps, un appariement exact entre les deux tables est réalisé sur l'ensemble des champs suivants : le premier nom, le premier prénom, le jour, le mois et l'année de naissance, le code officiel géographique (COG) de la commune de naissance et le sexe. Dans le cadre de l'appariement qualité, cette étape permet d'apparier environ 84% des 315 000 apprentis avec la source DSN. Une fois cet appariement réalisé, il reste à apparier environ 50 000 apprentis avec 7,2 millions de salariés ayant un contrat actif en décembre. Le volume de travail est ainsi déjà divisé par un facteur six[22]. Dans un second temps, l'union (sans doublons) des trois listes de paires suivantes est établie : les paires qui ont une distance faible entre les premiers noms et une distance faible entre les premiers prénoms et même département de naissance et même année de naissance ; les paires qui ont même date de naissance et même département de naissance ; les paires qui ont même premier nom et même premier prénom. La distance entre les noms et entre les prénoms utilisée dans InserJeunes est la distance de Levenshtein. Cette dernière est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'un nom/prénom à l'autre. L'union des différentes requêtes permet de bien couvrir tous les cas rencontrés fréquemment et ainsi de bien conserver presque toutes les paires « potentiellement intéressantes ». De plus, comme chaque requête est relativement précise, le nombre de paires « potentiellement intéressantes » retenues n'est pas trop élevé. Pour l'appariement qualité, cette méthode d'indexation conduit à retenir 1 million de paires soit un nombre de paires raisonnable qui pourra être traité suffisamment rapidement lors des étapes ultérieures du processus.

Ensuite, une similarité est calculée pour chacun des cinq couples d'identifiants indirects (ex : couples de noms, couples de date de naissance) de chaque paire. Chaque similarité est une mesure du degré de ressemblance des identifiants indirects considérés. Trois natures de similarités différentes sont utilisées dans InserJeunes selon la nature des variables mobilisées comme identifiants indirects. Tout d'abord, la similarité de Jaro-Winkler est utilisée pour les identifiants indirects noms et prénoms. Ensuite, une similarité ad hoc spécifique à InserJeunes a été élaborée pour les dates de naissance. Enfin, pour la variable sexe une similarité binaire est utilisée. Pour le code COG de la commune de naissance, la similarité est de 1 lorsque les codes sont identiques. S'ils sont différents, la similarité est de 0,5 si le code département est identique et de 0 sinon.

Quatrièmement, chaque paire est classifiée, c'est-à-dire que les paires supposées relever du même individu (i.e. lorsque les cinq similarités calculées à l'étape précédentes sont suffisamment élevées) sont acceptées et que les autres sont rejetées. Il n'y a donc pas de reprise manuelle lors de cette étape. Dans InserJeunes l'approche simple suivante a été retenue : on calcule une similarité globale pour chaque paire, fonction strictement croissante des similarités des différents champs ; les paires dont la similarité globale est supérieure à un certain seuil sont acceptées, les autres étant rejetées. La fonction et le seuil retenus sont choisis de manière empirique via l'analyse d'un échantillon de paires dont le statut (accepté ou rejeté) a été annoté manuellement. Cette méthode présente l'avantage de la simplicité mais le choix de la fonction et du seuil demeurent arbitraires donc rien ne garantit que ces choix soient optimaux. Nous avons cependant conservé cette méthode car les classifications que nous avons testées basées sur les forêts aléatoires ou les méthodologies de machines à vecteurs de support (support vector machine, SVM) n'ont pas donné de meilleurs résultats.

Comment expliquer ce résultat qui peut paraître a priori surprenant ? Une façon de représenter notre problème est de considérer que chaque paire est un point dans un espace à 5 dimensions, les dimensions étant similarités pour le nom, prénom, la date de naissance, commune de naissance et le sexe. La variable sexe étant très peu discriminante, elle pourrait être éliminée de l'analyse ce qui restreindrait l'espace à 4 dimensions. Dans chaque dimension les similarités prennent des valeurs entre 0 et 1.

[22] 315000/50000 ou 100/(100-84).

Le problème consiste donc à trouver une frontière de séparation entre les points paires acceptées et les points paires rejetées dans un espace [0 ;1]^4 soit un espace de toute petite taille. De plus, la zone de l'espace « proche » du point (1,1,1,1) correspond à la zone dans laquelle presque toutes les paires qu'on doit accepter se trouvent. Ainsi, les points correspondant aux paires qu'il faut accepter ne sont pas trop « mélangés » avec les points correspondant aux paires qu'il faut rejeter. Il est donc relativement facile de résoudre ce type de problème ce qui explique que toutes les méthodes testées donnent de manière équivalente de très bons résultats.

Enfin, la qualité du processus d'appariement est évaluée. Cela nécessite de disposer d'un échantillon de paires dont le statut (accepté ou rejeté) a été annoté manuellement et qui n'a pas été utilisé lors de l'étape de classification. Sur cet échantillon, la prédiction issue de la classification supervisée est comparée avec le véritable statut de la paire, c'est-à-dire celui établi manuellement, ce qui permet d'obtenir dans un premier temps quatre grandeurs : les vrais positifs (VP), les faux positifs (FP), les vrais négatifs (VN) et les faux négatifs (FN). Par exemple, une paire faux négatif est une paire rejetée par l'algorithme de classification mais acceptée par l'humain qui a réalisé l'annotation. A partir de ces quatre grandeurs il est possible d'établir plusieurs mesures de la qualité globale.

La mesure la plus connue est l'accuracy qui correspond à (VP+VN) / nombre total des paires. Mais cette mesure ainsi que toute mesure qui utilise les vrais négatifs n'est pas adaptée. Pourquoi ? Parce que les données sont déséquilibrées : il y a beaucoup de paires dont le vrai statut est rejeté et peu de paires dont le vrai statut est accepté. Dans le cas de l'appariement qualité environ 40 000 paires sont acceptées sur 1 million de paires donc au minimum 950 000 paires ont pour véritable statut « rejeté ». Un classifieur naïf qui rejette 100% des paires a donc une accuracy d'au moins (0+950 000) / (1 000 000) soit 95%.

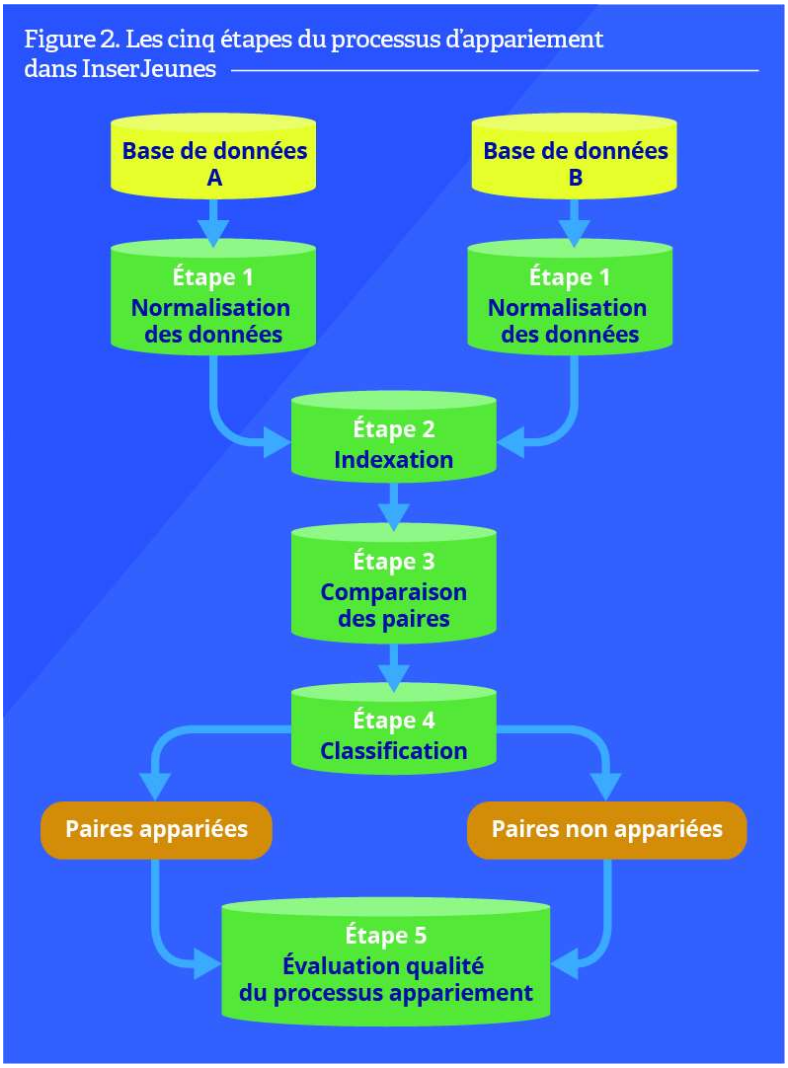
Trois mesures ont été retenues dans InserJeunes :

- la précision qui correspond à VP/(VP+FP). Par exemple, si la précision est de 80% alors cela veut dire que 80% des paires acceptées le sont à bon escient.

- le rappel qui correspond à VP/(VP+FN). Par exemple, si le rappel est de 90% alors cela veut dire que 90% des vraies paires ont été détectées par l'algorithme de classification.

- la f-mesure, moyenne harmonique de la précision et du rappel, qui correspond à 2*(précision* rappel) / (précision+ rappel).

Dans le cas de l'appariement qualité, la précision vaut 95% et le rappel 99%. Au total, 97% des apprentis sont appariés dans l'appariement qualité (84% via l'appariement direct et 13% via l'appariement approché) soit un taux d'appariement proche du taux d'appariement théorique de 100%.



3.3. Le développement d'un outil d'appariement spécifique

Plusieurs logiciels d'appariement ont été testés par l'équipe projet en charge d'InserJeunes. Seul l'outil MatchID répondait à nos besoins mais sa mise en œuvre s'est avérée relativement complexe. Suite à ce travail de benchmarking, il a été décidé de développer un outil d'appariement spécifique dans le cadre d'InserJeunes qui réponde aux quatre grands besoins détaillés ci-dessous.

Premièrement, les appariements doivent être rapides. L'outil d'appariement InserJeunes réalise l'appariement qualité en 15 mn (appariement d'environ 315 000 apprentis avec 7,5 millions de salariés). Cependant, pour des projets d'appariement sur des tables nettement plus volumineuses, l'outil développé pourrait soit nécessiter des adaptations soit s'avérer trop lent.

Deuxièmement, l'outil d'appariement doit être générique c'est-à-dire facilement adaptable pour tous les cas d'appariements sur identifiants indirects. Pour ce faire, la spécification de chaque appariement (par exemple : les champs comparés, la méthode de similarité choisie pour chaque champ, …) est décrite dans un fichier XML de spécification qui est ensuite interprété par l'outil. Cette façon de procéder permet également d'assurer une traçabilité complète de chaque appariement.

Troisièmement, étant donné que l'évaluation du processus d'appariement est fondamentale et que cela nécessite de disposer d'un échantillon de paires annotées manuellement, nous avons développé une interface d'annotation de paires ergonomique.

Quatrièmement, l'outil s'appuie sur plusieurs librairies open source ce qui a permis d'accélérer son développement et d'en faciliter la maintenance. L'indexation est réalisée en langage sql sur une base de données postgresql en mobilisant le module fuzzystrmatch, le calcul des similarités de Jaro-Winkler et de Levenshtein mobilise la librairie Python jellyfish et les algorithmes de machine learning sont réalisés avec la librairie Python scikit learn.

L'outil d'appariement d'InserJeunes sera prochainement mis à disposition en open source par la DEPP.

4. Le taux d'emploi par établissement accompagné d'un calcul de valeur ajoutée via des modèles économétriques multiniveaux

Comme indiqué dans la partie 1, le taux d'emploi donne la part, parmi les élèves ne poursuivant pas leurs études, de ceux qui disposent d'un contrat de travail dans le secteur privé, sur une semaine de référence, 6 mois (et à terme 12, 18 ou 24 mois) suivant la fin de leurs études. Cet indicateur vise à mesurer l'insertion en emploi des sortants de formation à différentes dates de façon à suivre leur trajectoire en début de carrière professionnelle.

Les taux d'emploi calculés à partir du dispositif Inserjeunes ne sont pas comparables directement aux estimations de taux d'emploi réalisées précédemment à l'aide des enquêtes sur l'insertion professionnelles des lycéens et des apprentis (enquêtes IVA et IPA). La transition vers le nouveau dispositif provoque une rupture de série, les taux d'emploi calculés par InserJeunes se situant en moyenne 10 points plus bas. Cette baisse s'explique par deux effets principaux. Premièrement, un changement de la période de référence entre les deux dispositifs : InserJeunes mesure l'insertion professionnelle sur une semaine de référence en janvier n+1 alors que dans les enquêtes on mesurait cette dernière sur le mois de février n+1, ce mois d'écart entre les deux dispositifs conduit donc automatiquement à une baisse du taux d'emploi estimé à environ 6 points en mesurant le taux d'emploi sur ces deux périodes de référence à partir de la DSN. Deuxièmement, à ce stade, l'emploi est mesuré sur le champ du salariat privé en France. Il ne mesure donc pas l'emploi à l'étranger, l'emploi non salarié, l'emploi public, l'emploi auprès de particuliers employeurs ou à l'aide des titres emploi simplifié agricole (TESA). L'emploi mesuré via la DSN peut donc être minoré dans certaines formations par rapport à l'insertion professionnelle des jeunes sortants de voie professionnelle. Ce champ non couvert par la DSN représente environ 2% de l'emploi des sortants de voie professionnelle scolaire et 4% de l'emploi des sortants d'apprentissage selon les estimations réalisées à partir des enquêtes IVA/IPA. Les établissements qui préparent à des spécialités conduisant principalement à des emplois dans la fonction publique ou dans les autres secteurs cités ci-dessus ne sont donc pas couverts par Inserjeunes à ce stade. A terme, le champ de la DSN permettra de couvrir l'ensemble du champ salarié (public et privé, y compris auprès des particuliers employeurs ou réalisés à l'aide du TESA).

Le reste de la baisse s'explique par des protocoles différents. En particulier, dans les enquêtes, tous les jeunes ne répondaient pas aux enquêtes ce qui pouvait s'accompagner d'un biais sur le profil des répondants en particulier en lien avec leur situation sur le marché de l'emploi. Cette non-réponse pouvait donc induire un biais sur le taux d'emploi estimé, les jeunes sans emploi étant sans doute moins enclins à répondre à l'enquête. Le nouveau dispositif, reposant sur des données administratives, n'est pas affecté par ce biais.

Une simple comparaison des taux d'emploi calculés entre établissements ne suffit pas. En effet, l'insertion dépend en partie de facteurs extérieurs à l'établissement : profil des jeunes, type et niveau de formation, spécialité de formation et marché du travail local. Par exemple, comparer directement le taux d'emploi d'un établissement n'offrant que des formations en CAP avec celui d'un établissement n'offrant que des formations en BTS est peu pertinent. Ainsi, pour compléter l'information sur l'insertion professionnelle, la loi du 5 septembre 2018 pour la liberté de choisir son avenir professionnel prévoit également la publication d'un indicateur de « valeur ajoutée », qui compare le taux d'emploi des élèves sortants de cet établissement au taux d'emploi « attendu »,

calculé comme le taux d'emploi moyen d'établissements similaires (en termes de profil des jeunes, type et niveau de formation, spécialité de formation et marché du travail local). La valeur ajoutée permet donc de tenir compte de l'ensemble de ces caractéristiques pour apprécier l'insertion professionnelle des jeunes en neutralisant ce qui ne tient pas à l'établissement lui-même[23] (voir [9]).

La valeur ajoutée est égale à la différence entre le taux d'emploi observé à 6 mois de l'établissement (le taux réel) et le taux d'emploi attendu à 6 mois (le taux issu d'un modèle statistique) :

Valeur ajoutée de l'établissement sur le taux d'emploi à 6 mois	=	Taux d'emploi observé à 6 mois de l'établissement	-	Taux d'emploi attendu à 6 mois de l'établissement

Le taux d'emploi observé à 6 mois est le taux d'emploi mesuré par InserJeunes via l'appariement des sortants de formation avec les contrats des salariés issus de la Déclaration Sociale Nominative (DSN).

Le taux d'emploi attendu à 6 mois est la moyenne, au niveau de l'établissement, des probabilités estimées par des modèles multiniveaux que les élèves de cet établissement soient en emploi à 6 mois. Il s'interprète comme le taux d'emploi moyen des élèves accueillis dans des établissements comparables en termes de profil des élèves, de formations dispensées et domiciliés dans une zone d'emploi au taux de chômage comparable.

La valeur ajoutée, différence entre deux taux, s'exprime en point de pourcentage et peut-être positive ou négative. Par exemple, un établissement avec un taux d'emploi observé à 65 % et un taux d'emploi attendu à 60 % aura une valeur ajoutée de +5.

Dans la suite de cette partie, ne sont repris que les principaux éléments du mode de calcul du taux d'emploi attendu et de la constitution des modèles. Le détail est disponible dans le document de travail série « Méthodes » de Midy, L. et Deschamps G. (voir [9]).

4.1. Le mode de calcul du taux d'emploi attendu à 6 mois de l'établissement

Pour calculer le taux d'emploi attendu à 6 mois de l'établissement, on utilise des modèles économétriques Logit multiniveaux à effet aléatoire à 3 niveaux : élève, établissement et zone d'emploi de la commune de résidence de l'élève. Les niveaux établissement et zone d'emploi ne sont pas emboîtés (voir [7]), la commune de l'établissement fréquenté par un jeune n'étant pas forcément dans la zone d'emploi de sa commune de résidence.

Plus précisément, le modèle est le suivant. Pour chaque jeune i appartenant à un établissement j et dans une zone d'emploi k, la probabilité d'avoir un emploi P_{ijk}, pour un jeune i scolarisé dans un établissement j, résidant dans une zone d'emploi k, est modélisée de la façon suivante :

$$\text{Logit}(P_{ijk}) = \beta_0 + \beta_1 X_{ijk} + \beta_2 W_j + \beta_3 Z_k + \alpha_j + \gamma_k$$

Avec :

- β_0 est la constante du modèle.
- X_{ijk} représente l'ensemble des variables individuelles du modèle (sexe, âge, diplôme, catégorie socioprofessionnelle des parents, obtention du diplôme…).
- W_j correspond à la part d'élèves en situation de handicap au sein de l'établissement j. Cette variable est présente uniquement pour les modèles des lycées professionnels sous tutelle du ministère de l'éducation nationale.

[23] Une méthodologie similaire est utilisée dans le cadre de la diffusion des indicateurs produits par la Depp depuis de nombreuses années pour les lycées, les indicateurs de valeur ajoutée des lycées (IVAL). Les IVAL mesurent la valeur ajoutée des établissements sur le taux de réussite au baccalauréat, le taux d'accès au baccalauréat et le taux de mentions au baccalauréat. Des lycées prestigieux ayant des taux de réussite au bac de 100 % peuvent ainsi avoir une valeur ajoutée faible et inversement.

- Z_k correspond au taux de chômage de la zone d'emploi.
- α_j est l'erreur aléatoire du modèle (le résidu) pour l'établissement j. Cette erreur est supposée suivre une loi normale de moyenne nulle.
- γ_k est l'erreur aléatoire du modèle (le résidu) pour la zone d'emploi k. Cette erreur est supposée suivre une loi normale de moyenne nulle.

Comme le niveau établissement n'est pas strictement emboîté dans le niveau zone d'emploi, le résidu de niveau établissement s'écrit α_j et ne dépend pas de k.
Puis on calcule le taux d'emploi attendu à 6 mois de l'établissement qui correspond à la moyenne des probabilités individuelles d'être en emploi après avoir soustrait le résidu établissement α_j.

4.2. Élaboration des 9 modèles par ajout successif de variables

Les niveaux de diplômes ainsi que la filière de formation (par apprentissage ou par voie scolaire) étant déterminants pour l'insertion professionnelle des sortants, des modèles par filière et par niveau de diplôme (CAP, Baccalauréat professionnel, BTS ainsi que Brevet Professionnel pour la filière apprentissage) ont été privilégiés. Un modèle spécifique pour les apprentis qui sont en formation dans les lycées a également été estimé. En effet, ces apprentis sont comptés deux fois dans les indicateurs : une fois en tant qu'apprentis liés à leur CFA et une fois dans les lycées professionnels dans lesquels ils sont accueillis. Ainsi pour un lycée professionnel ayant des apprentis nous diffusons la valeur ajoutée de l'ensemble de l'établissement, ainsi que la valeur ajoutée des deux populations qui le composent : les lycéens et les apprentis.

Au total, 9 modèles ont ainsi été estimés pour calculer la probabilité d'être en emploi des élèves d'InserJeunes : 4 modèles pour les sortants d'apprentissage, 1 modèle pour les sortants d'apprentissage en formation dans les lycées, 3 modèles pour les sortants de voie professionnelle scolaire d'établissements sous tutelle du ministère de l'éducation nationale et 1 modèle pour les sortants de voie professionnelle scolaire d'établissements sous tutelle du ministère de l'agriculture.

Pour les sortants d'apprentissage en lycée ainsi que pour les sortants de voie professionnelle scolaire sous tutelle du ministère de l'agriculture, les effectifs n'étaient pas suffisants pour faire des modèles par niveau de formation, par conséquent la variable niveau de diplôme est introduite dans les variables explicatives de ces deux modèles et a un fort pouvoir explicatif.

Pour chacun des modèles, on cherche à obtenir un modèle **parcimonieux** qui ne contient donc pas toutes les variables disponibles pour la modélisation mais qui explique une part importante de la variance établissement. En effet, plus la variance de niveau établissement est faible, plus les variables intégrées dans le modèle expliquent « bien » le fait que l'élève soit en emploi ou non, et moins la valeur ajoutée sera dispersée entre les établissements.

L'élaboration de modèles de façon imbriquée permet de facilement comparer les modèles les uns avec les autres en regardant la déviance ce qui n'est possible que dans le cas où un modèle est inclus dans l'autre et de maîtriser l'impact et la significativité de chaque variable ajoutée. Dans le cadre des modèles multiniveaux, nous commençons par un modèle vide (sans variables explicatives) avec constante aléatoire pour le niveau établissement. Commencer par un modèle sans variables explicatives permet de s'assurer de la présence de l'effet des niveaux : il s'agit alors d'une simple décomposition de la variance en variance inter-classes (variance entre établissements) et variance intra-classes (variance entre individus). Or, la variance individuelle doit être fixée pour rendre un modèle multiniveaux binaire identifiable. Dans le cas des modèles logistiques, elle est fixée à $\pi^2/3$, ce qui correspond à la variance d'une loi Logit standard et elle ne varie donc pas lorsqu'on introduit de nouvelles variables

Nous ajoutons ensuite successivement les variables de niveau individuel, les variables de niveau établissement et la variable zone d'emploi. L'ajout de chaque variable est évalué en regardant les différents critères listés ci-dessous :

- Baisse de la déviance et sa significativité : La déviance correspond à la valeur -2 * Log (L) ou L est la vraisemblance du modèle. Plus la déviance diminue par rapport au modèle de référence, mieux le modèle décrit les données. Par définition, la déviance diminue avec le nombre de paramètres ajoutés au modèle et ne pénalise pas les modèles trop complexes. Par contre, comme la diminution de la déviance d'un modèle à l'autre suit une loi du Chi2 il est possible de tester la significativité de l'ajout successif de chaque variable. Le nombre de degrés de liberté de la loi du Chi2 est déterminé ici par le nombre de paramètres supplémentaires à estimer dans le modèle le plus complexe.

- Baisse du critère d'information d'Akaike (AIC) : L'AIC prend en compte la diminution de la déviance mais est pénalisé par deux fois le nombre k de paramètres ajoutés, soit AIC = - 2 * log(L) + 2 * k. L'AIC représente donc un compromis entre le biais (qui diminue avec le nombre de paramètres) et la parcimonie (nécessité de décrire les données avec le plus petit nombre de paramètres possible). En choisissant le modèle avec l'AIC le plus faible, on s'assure que le modèle est robuste et ne contient que les variables ayant un pouvoir explicatif important.

- Diminution de la variance de niveau établissement : Les modèles multiniveaux permettent d'estimer la variance des différents niveaux et leur part dans la variance totale du modèle. Par rapport au modèle sans aucune variable explicative, plus on ajoute de variables individuelles et plus la variance de niveau établissement diminue.

- Dispersion de la valeur ajoutée : Nous regardons d'abord l'écart type de la valeur ajoutée, plus il est faible et moins la valeur ajoutée est dispersée. Puis en analysant la répartition de la valeur ajoutée en valeur absolue par classe, nous nous attendons à voir de moins en moins d'établissements avec des valeurs ajoutées extrêmes.

- D de Somers : Le D de Somers est un indicateur basé sur les rangs. On considère toutes les paires d'observations ayant des valeurs observées de Y différentes, soient 1 et 0 et on les répartit dans 3 groupes :

    - les paires concordantes : celles pour lesquelles l'observation où Y = 1 a une probabilité estimée que Y = 1 plus grande que l'observation où Y = 0.

    - les paires discordantes : celles pour lesquelles l'observation où Y = 1 a une probabilité estimée que Y = 1 plus faible que l'observation où Y = 0.

    - les paires « ex-aequo » : celles pour lesquelles l'observation où Y = 1 a une probabilité estimée que Y = 1 égale à celle de l'observation où Y = 0.

Le D de Somers qui est défini par $\frac{nombre\ de\ paires\ concordantes - nombre\ de\ paires\ discordantes}{nombre\ de\ paires\ ayant\ des\ valeurs\ observées\ de\ Y\ différentes}$ varie dans [- 1, + 1], il est égal à 0 quand il n'y a pas d'association, il tend vers 1 lorsque l'association est très forte et positive et vers -1 lorsque l'association est très forte et négative.

- Examen de la significativité des coefficients ajoutés : Nous conservons les variables quantitatives qui sont significatives dans au moins un des modèles par niveau de diplôme. De même, nous conservons les variables qualitatives lorsqu'au moins une des modalités est significative dans un des modèles par niveau de diplôme.

- Examen des Odds Ratio : Les Odds Ratio donnent une information sur l'ampleur de la relation entre une variable explicative et la variable d'intérêt. L'examen des Odds Ratio permet de

s'assurer que les variables que l'on garde dans les modèles ont un impact important sur la variable d'intérêt.

L'ensemble de ces différents indicateurs sont utilisés pour juger du degré de pertinence du modèle m+1 emboité dans le modèle m. La plupart du temps tous les indicateurs donnent des résultats cohérents (ex : baisse significative de la déviance et augmentation du D de Somers) et lorsque ce n'est pas le cas la majorité des indicateurs restent cohérents entre eux.

4.3. Principaux résultats

Le choix des 9 modèles a été réalisé en regardant les différents critères indiqués précédemment. La variance entre établissements diminue entre les modèles vides (sans aucune variable explicative) et les modèles complets (elle est au minimum divisée par deux).

Les variables retenues dans les modèles sont : un regroupement de spécialités de formation, un regroupement de codes de la nomenclature d'activités française (NAF) de l'établissement d'apprentissage (donc cette variable est spécifique au champ apprentissage), un regroupement de la variable âge en classes, une indicatrice prenant la valeur 1 si le diplôme correspond à une mention complémentaire pour les modèles pouvant avoir ce type de diplôme, le sexe de l'élève, la nomenclature des professions et catégories socioprofessionnelles (PCS) en une position du responsable de l'élève, la situation avant l'apprentissage (uniquement pour les apprentis), une variable combinant l'obtention du diplôme et le résultat à l'examen, une variable sur le handicap, le taux de chômage annuel au niveau de la zone d'emploi de la commune de résidence de l'élève. En ce qui concerne la variable sur le handicap, la reconnaissance de la qualité de travailleur handicapé (RQTH) pour les sortants d'apprentissage est introduite au niveau individuel, Pour les modèles sur les jeunes sortant pour la voie professionnelle scolaire en lycée, la part d'élèves en situation de handicap a été introduite, cette variable n'étant pas disponible au niveau individuel. L'annexe 1 donne une présentation détaillée de ces variables.

Le regroupement de spécialités de formation et le regroupement de NAF pour les CFA sont les variables qui ont le plus fort impact sur les modèles : diminution de la déviance, de l'AIC, de la variance établissement et de la dispersion de la valeur ajoutée. L'ajout au niveau établissement de variables agrégeant des données individuelles (part des élèves de l'établissement selon l'âge, selon les formations) est significatif mais a été abandonné pour limiter la complexité du modèle. L'ajout du taux de chômage de la zone d'emploi est significatif dans tous les modèles. La variable handicap a globalement peu d'impact dans les modèles mais à un très fort impact pour un petit nombre d'établissements accueillant majoritairement voire exclusivement des élèves/apprentis en situation de handicap.

Les résultats détaillés des différents modèles sont disponibles en annexe 1 (hors celui portant sur les sortants de voie professionnelle scolaire d'établissements sous tutelle du ministère de l'agriculture).

5. Le respect du RGPD, les bonnes pratiques et innovations mises en œuvre dans InserJeunes.

Le système d'information InserJeunes comporte de nombreuses données à caractère personnel (DCP) dont les noms et prénoms des élèves et apprentis (DCP directement identifiantes) et l'INE (DCP indirectement identifiante). Il faut donc veiller à respecter le cadre du RGPD notamment « pseudonymiser » les INE, supprimer les DCP dès qu'elles ne sont plus nécessaires aux traitements et sécuriser le système d'information InserJeunes ainsi que les échanges de données entre administrations.

5.1. La suppression des données à caractère personnel et la pseudonymisation des INE dans InserJeunes

Les appariements sur identifiants indirects (voir figure 2) sont réalisés en utilisant les noms, prénoms, sexe, date et commune de naissance soit des données à caractère personnel (DCP) présents dans les différents fichiers. L'usage de DCP étant encadré par le règlement général sur la protection des données (RGDP), le dispositif InserJeunes a fait l'objet d'une déclaration au registre des traitements suivis par le délégué à la protection des données (DPD) du ministère. Une analyse d'impact relative à la protection des données (AIPD) a également été réalisée. En effet, la réalisation d'une AIPD est obligatoire d'une part pour certains types de traitements (cette liste ayant été établie par la CNIL[24]) et d'autre part lorsque au moins 2 critères parmi une liste de 9 critères s'applique au traitement ce qui est le cas d'InserJeunes qui remplit les 3 critères suivants : collecte de données personnelles à large échelle, croisement de données et personnes vulnérables (patients, personnes âgées, enfants, etc.). Schématiquement, une AIPD comporte trois parties : une description du traitement mis en œuvre, l'évaluation de la nécessité et de la proportionnalité de collecte de DCP, une analyse des risques de sécurité ainsi que leur impact potentiel sur la vie privée. Le RGPD impose de limiter au strict nécessaire, compte tenu des finalités du traitement, la collecte et la conservation de DCP et de les protéger en sécurisant correctement le système d'information.

Le RGPD promeut la « pseudonymisation » des données à caractère personnel lorsque cela peut être pertinent[25]. La « pseudonymisation » est définie dans l'article 4 du RGPD : il s'agit du traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles afin de garantir que les données à caractère personnel ne sont pas attribuées à une personne physique identifiée ou identifiable.

Une façon de procéder[26] consiste à recourir à une fonction dite de « ***hachage*** », qui présente la particularité, par rapport aux algorithmes de chiffrement standards, d'être irréversible : il n'est donc pas possible par exemple de retrouver l'INE à partir du seul INE « pseudonymisé » par hachage, même si l'on connaît la fonction de hachage utilisée. Toutefois, en dépit de cette irréversibilité de principe, cette technique peut être mise en échec en reconstituant, par réitération, une table de correspondance. Cette méthode pour casser l'anonymisation suppose d'importants moyens informatiques : elle consiste à appliquer la fonction de hachage à l'ensemble des identifiants possibles (par exemple, l'ensemble des noms et prénoms des individus susceptibles d'appartenir à la base de données). Ainsi, on retrouve, pour chacun, le pseudonyme unique qui lui est attribué par la fonction de hachage initialement utilisée. La sécurité de l'anonymisation peut être renforcée en ajoutant préalablement aux identifiants initiaux une clé secrète arbitraire (également dénommée « sel ») : par exemple au nom « *Jean Dupont* », on associe la clé « *azerty* », pour donner un second identifiant « *Jean Dupontazerty* », qu'on soumet alors à la fonction de hachage. Ainsi pour reconstituer la table de correspondance, il faudra donc non plus seulement tester l'ensemble des noms et prénoms possibles, ce qui est relativement facile, mais aussi l'ensemble des modifications que ces identifiants sont susceptibles de connaître à partir de clés inconnues. La sécurité du dispositif repose cependant encore une fois sur la confidentialité des outils utilisés : la clé secrète arbitraire et la fonction de hachage utilisée. Il est encore possible de durcir l'anonymisation, en procédant à un double hachage avec clé secrète, qui consiste à réaliser une première fois l'opération, et à soumettre le pseudonyme obtenu à une seconde fonction de hachage avec clé secrète.

[24] https://www.cnil.fr/sites/default/files/atoms/files/liste-traitements-aipd-requise.pdf

[25] Notamment dans l'Article 25 - Protection des données dès la conception et protection des données par défaut *« le responsable du traitement met en œuvre, tant au moment de la détermination des moyens du traitement qu'au moment du traitement lui-même, des mesures techniques et organisationnelles appropriées, telles que la pseudonymisation, qui sont destinées à mettre en œuvre les principes relatifs à la protection des données ».*

[26] https://www.senat.fr/rap/r13-469/r13-4697.html

Dans InserJeunes, certaines DCP sont supprimées automatiquement au fur et à mesure du déroulement du processus statistique ce qui permet de garantir leur effacement dès que ces DCP ne sont plus nécessaires pour réaliser les traitements InserJeunes. Par exemple, les fichiers fournis par les producteurs sont supprimés dès qu'ils ont été intégrés correctement dans le SI InserJeunes. De même, les noms et prénoms des élèves qui sont en dehors du champ InserJeunes sont supprimés dès que le champ est établi. Les autres DCP sont supprimées via des traitements batchs spécifiques lancés par le responsable du traitement via une application de pilotage d'InserJeunes.

En ce qui concerne l'INE, une méthode de double hachage a été mise en place de la façon suivante :

- pour chaque nouvel INE, on crée une clé secrète aléatoire et on stocke la table de passage INE/clé secrète. Lorsqu'on doit hacher un INE déjà existant dans la table de passage, on récupère la clé secrète dans cette dernière.

- la chaîne résultant de la concaténation de l'INE et de sa clé secrète est ensuite hachée avec l'algorithme de hachage SHA512 figurant dans la liste des algorithmes reconnus et jugés sûrs par la CNIL[27].

- puis on « hache » la chaine résultant de la concaténation du hash obtenu à l'étape précédente et d'une clé secrète unique (donc commune à tous les INE). On procède donc à 2 hachages successifs et l'INE « pseudonymisé » est le résultat obtenu à l'issu de ces 2 hachages.

5.2. L'organisation des données dans InserJeunes, leur sécurisation et leur traçabilité

Le processus de traitement des données dans InserJeunes peut être vu comme comportant 3 niveaux. Les sources fournies par les producteurs sont intégrées sans modification de manière brute ce qui donne des données que l'on peut qualifier de niveau « bronze » d'après la segmentation des données recommandée par la société Databricks spécialisée dans la mise en œuvre de lacs de données dans un contexte big data[28]. Elle contribue à assurer une traçabilité complète des données, chaque état de chaque variable étant conservé dans le système d'information.

Ensuite, nous réalisons des contrôles des sources intégrées par rapport à des nomenclatures et des référentiels existants, des normalisations (par exemple, normalisation des champs nom et prénom), des mises en cohérence (par exemple, supprimer les doublons entre les différentes sources comme le cas d'un apprenti en formation dans un lycée professionnel qui peut être présent à la fois dans la source recensant l'ensemble des apprentis[29] et dans la source recensant les élèves dans les lycées[30]), des appariements sur identifiants directs et sur identifiants indirects et enfin des enrichissements (par exemple, ajout de la probabilité d'être en emploi via des modèles économétriques multiniveaux). A l'issue de ces traitements, nous obtenons des données de niveau « argent ». Enfin, nous calculons des données agrégées non soumise au secret, soit des données de niveau « or », qui sont ensuite diffusées.

Dans InserJeunes, un niveau élevé de traçabilité sur les traitements est mis en œuvre car à chaque traitement batch réalisé est associé un fichier de type « journal de bord » (la « log ») qui détaille et liste tous les contrôles et actions réalisées sur les données durant ce traitement. Ce fichier est précieux pour aider l'équipe statistique en charge de la production à voir si le traitement s'est ou non bien déroulé.

[27] https://www.cnil.fr/fr/securite-chiffrer-garantir-lintegrite-ou-signer
[28] https://databricks.com/blog/2019/08/14/productionizing-machine-learning-with-delta-lake.html
[29] Source SIFA.
[30] Source SYSCA SCO.

Enfin, afin d'augmenter le niveau de sécurité d'InserJeunes, le système a été découpé en 3 parties cloisonnées :

- la partie « réception/traitement » dans laquelle on stocke toutes les données individuelles et les DCP. Tout le processus statistique se déroule dans cet environnement qui est très sécurisé à accès très restreint, réservé uniquement aux personnes en charge de la production des indicateurs InserJeunes à la Depp.

- la partie « diffusion » dans laquelle on stocke uniquement des données agrégées qui sont donc sans secret. Cette partie alimente notamment le site de consultation grand public.

- la partie « étude » avec des bases à destination des chargés d'étude habilités au sein de la Depp et de la Dares, ainsi qu'à des chercheurs dans le cadre de conventions : des données individuelles sont présentes dans ces bases, mais elles ne comportent plus de données à caractère personnel. L'INE « pseudonymisé » est également à accès restreint et dépend du destinataire de la base.

Tous les développements InserJeunes ont fait l'objet d'un audit de sécurité et tous les échanges de données entre la Depp et la Dares passent par le réseau interministériel de l'État (RIE) où le flux est chiffré ainsi que les données elles-mêmes.

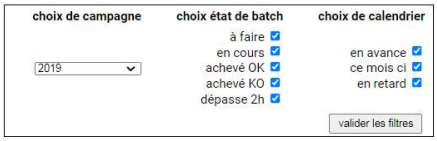
5.3. L'application de pilotage de lancement des programmes

L'équipe statistique en charge de la production statistique dispose d'une application de pilotage de la production InserJeunes qui lui permet de lancer en autonomie tous les programmes en production (voir figure 3). Nous avons mis en place la gestion de toutes les dépendances entre programmes sous la forme d'un graphe orienté acyclique[31] développé de manière ad hoc[32]. Concrètement cela signifie qu'un programme n'est lançable que lorsque tous ses prérequis sont remplis (présence de fichier(s) ou table(s), programme(s) prédécesseur(s) achevés correctement). De même, Il est possible de relancer n'importe quel programme mais cela impose de relancer ensuite tous les programmes successeurs du programme qu'on vient de relancer.

Figure 3. Copie d'écran de l'application de pilotage des programmes

[31] En anglais directed acyclic graph ou DAG.

[32] Il existe des logiciels open source d'orchestration de batchs implémentant les DAG comme apache airflow mais nous avons préféré refaire un développement ad hoc car notre besoin était relativement simple et nous voulions proposer une application de pilotage unique comportant également d'autres fonctionnalités de gestion (ex : gestion de comptes).



campagne sortants été	libellé du batch	nom 1er paramètre	valeur 1er paramètre	nom 2nd paramètre	valeur 2nd paramètre	état du batch	calendrier	lancer le batch
2019	chargement_referentiel					achevé OK		lancer le batch
afficher/cacher les prérequis du batch								
2019	chargement_scolege	millesime	2019_2020			achevé OK		lancer le batch
afficher/cacher les prérequis du batch								
2019	chargement_sysca_sco	millesime	2019_2020			achevé OK		lancer le batch
afficher/cacher les prérequis du batch								
le fichier SYSCA_SCO_2019_2020.csv ou la table SYSCA_SCO_brut est disponible le batch chargement_referentiel,None,None,None,None (campagne 2019) est achevé OK								
2019	chargement_voeu_ifsi	millesime	2019_2020			achevé OK		lancer le batch
afficher/cacher les prérequis du batch								
2019	chargement_dger	millesime	2019_2020			achevé OK		lancer le batch
afficher/cacher les prérequis du batch								
2019	chargement_sismmo	date_debut	01122019	date_fin	31012020	achevé OK		lancer le batch
afficher/cacher les prérequis du batch								
le fichier SISMMO_CONTRAT_01122019_31012020.csv ou la table SISMMO_CONTRAT_brut_01122019_31012020 est disponible le fichier SISMMO_SALARIE_01122019_31012020.csv ou la table SISMMO_SALARIE_brut_01122019_31012020 est disponible le batch chargement_referentiel,None,None,None,None (campagne 2019) est achevé OK								
2019	chargement_examens	millesime	2018_2019			achevé OK		lancer le batch
afficher/cacher les prérequis du batch								

Commentaire : le batch de chargement de la source « sysca sco » pour l'année scolaire 2019-2020 a pour prérequis la présence du fichier CSV de données (ou de la table brute en BDD si le chargement a déjà été fait) et la bonne exécution du batch de changement du référentiel. Ces deux prérequis étant remplis le batch est lançable via l'IHM.

6. Conclusion

Le système d'information InserJeunes sous co-maitrise d'ouvrage Depp/Dares permet en particulier d'avoir des taux d'emploi des jeunes sortant d'apprentissage ou de voie scolaire pour les niveaux 3 à 5 à un niveau de finesse très détaillé : par établissement et type de diplôme. Si les effectifs sont suffisants, on peut même disposer du taux d'emploi au niveau du diplôme fin au sein d'un établissement donné. Le développement de ce SI a duré presque trois années. Il a bénéficié d'un financement du Fonds pour la Transformation de l'Action Publique (FTAP) et dans ce cadre les différents jalons imposés dans le cahier des charges ont été respectés dans les délais prévus.

Plusieurs défis méthodologiques ont été relevés durant la phase projet avec en particulier le développement d'un outil d'appariement ad hoc et la mise au point d'une valeur ajoutée accompagnant l'indicateur de taux d'emploi à 6 mois qui permet d'estimer l'apport propre de chaque centre de formation d'apprentis et de chaque lycée professionnel. De plus, la manipulation de nombreuses données à caractère personnel dont en particulier les noms et prénoms des élèves et apprentis a nécessité une vigilance particulière sur le respect du RGPD en supprimant ces données dès qu'elles ne sont plus nécessaires aux traitements et en les sécurisant au mieux dans les traitements et leur stockage. Les différentes parties décrites dans cet article peuvent être utiles dans tout type d'appariement sur identifiants indirects utilisant noms, prénoms, sexe, date et lieu de naissance, et plus particulièrement dans un système d'information qui « poursuivrait » le champ d'InserJeunes en calculant des taux d'emploi pour les jeunes sortant de l'enseignement supérieur.

Les premiers indicateurs issus de ce nouveau système d'information au niveau des établissements et au niveau national ont été diffusés en février 2021 d'une part au niveau national par diplôme fin sur les jeunes sortis du système éducatif en 2019 et d'autre part au niveau des établissements et type de diplôme sur le cumul des sortants 2019 et 2018. Cette diffusion revêt plusieurs formes dont en particulier la publication de deux études (voir [3] et [4]) et la mise à disposition de fichiers Excel sur le site du ministère en charge de l'Education nationale ainsi que sur le site du ministère de l'emploi. Un site de consultation dédié a également été développé pour présenter les principaux résultats d'InserJeunes au grand public https://www.inserjeunes.education.gouv.fr/diffusion/accueil. La première diffusion a nécessité de l'accompagnement auprès des établissements, lycées et centres de

formation d'apprentis pour expliciter chacun des indicateurs diffusés ainsi que la notion de valeur ajoutée. En mai 2021, le taux d'emploi 12 mois après la sortie de ces jeunes a également été publié au niveau national et au niveau établissement (voir [2]). En régime courant, plusieurs diffusions sont prévues par an : une première diffusion qui concerne le taux d'emploi à 6 mois des sortants en n, une seconde avec le taux d'emploi à 12 mois des sortants en n et une concernant l'ensemble des taux d'emploi à 6, 12, 18 et 24 mois des sortants en n-1.

Bibliographie

[1] Christen P., 2012, data matching concepts and technique for record linkage, entity resolution and duplicate detection, Springer

[2] Collin, C., Marchal N., 2021, "Des lycéens professionnels et des apprentis mieux insérés 12 mois après leur sortie d'études en juillet 2020 que 6 mois après, malgré la crise", Note d'information n°21.24, mai, DEPP-MENJS

[3] Collin, C., Marchal N., 2021, "Six mois après leur sortie en 2019 du système éducatif, 41 % des lycéens professionnels sont en emploi salarié", Note d'information, n°21.06, février, DEPP-MENJS

[4] Collin, C., Marchal N., 2021, "Six mois après leur sortie en 2019 du système éducatif, 62 % des apprentis de niveau CAP à BTS sont en emploi salarié", Note d'information, n°21.07, février, DEPP-MENJS

[5] Evain, F., 2020, « Indicateurs de valeur ajoutée des lycées : du pilotage interne à la diffusion grand public », Courrier des statistiques N5 – 2020, Insee

[6] Evain, F., Evrard L., 2017, « Une meilleure mesure de la performance des lycées », refonte de la méthodologie des IVAL (session 2015), Education & Formations, n°94, 2017

[7] Givord, P., Guillerm, M., 2016, « Les modèles multiniveaux », document « Méthodologie Statistique », n° M2016/05, Insee

[8] Midy, L., 2021, « InserJeunes – taux d'interruption en cours de formation », document de travail série « Méthodes » n°2021-M02, DEPP-MENJS

[9] Midy, L., Deschamps G., 2021, « InserJeunes – calcul de valeur ajoutée », document de travail série « Méthodes » n°2021-M01, DEPP-MENJS

[10] Midy, L., 2021, « un outil d'appariement sur identifiants indirects : l'exemple du système d'information sur l'insertion des jeunes », Courrier des statistiques N6 – 2021, Insee

Annexe 1 : Détails des modèles finaux pour les sortants 2018

Au total, 9 modèles ont été estimés : 4 modèles pour les sortants d'apprentissage (1er modèle : CAP, mentions complémentaires et autres formations de niveau 3, 2ème modèle : Brevet professionnel (BP), 3ème modèle : Bac pro, mentions complémentaires et autres formations de niveau 4, 4ème modèle : BTS et autres formations de niveau 5), 1 modèle pour les sortants d'apprentissage en formation dans les lycées, 3 modèles pour les sortants de voie professionnelle scolaire d'établissements sous tutelle du ministère de l'éducation nationale (1er modèle : CAP, mentions complémentaires et autres formations de niveau 3, 2ème modèle : Bac pro, mentions complémentaires et autres formations de niveau 4, 4ème modèle : BTS et autres formations de niveau 5) et 1 modèle pour les sortants de voie professionnelle scolaire d'établissements sous tutelle du ministère de l'agriculture. Ce dernier modèle n'est pas présenté dans cette annexe.

Indicateurs d'ajustement des modèles

		Apprentissage				Apprentis en lycée	Voie professionnelle scolaire		
Niveau de diplôme		niveau 3	niveau 4 BP	niveau 4 autres	niveau 5	tous niveaux	niveau 3	niveau 4	niveau 5
Taux d'insertion 6 mois après la sortie		54%	73%	64%	69%	65%	26%	37%	55%
Nombre d'élèves		41 248	14 256	18 584	28 368	12 591	26 213	75 374	56 444
Nombre d'établissements		619	415	601	629	721	1 637	2 046	1 734
Nombre d'élèves (établissements avec effecti		40 986	14 127	18 360	27 683	12 412	24 978	73 296	52 698
Nombre d'établissements (avec effectif >=20)		581	396	567	551	683	1 487	1 830	1 416
Critères ajustement	deviance (-2 log vraissembla	53 285	15 440	22 493	33 057	14 869	26 819	95 103	76 143
	AIC	53 349	15 502	22 557	33 119	14 939	26 869	95 153	76 191
Estimation variance	variance établissement	0,05	0,04	0,05	0,04	0,06	0,11	0,06	0,07
	variance zone emploi	0,02	0,04	0,02	0,04	0,01	0,04	0,03	0,02
	total	0,07	0,08	0,07	0,08	0,07	0,14	0,09	0,09
filtre 20 sortants minimum par établissement									
Valeur Ajoutée	écart type VA	13,3	14,6	14,9	12,7	20	16	11,1	14,5
(étab avec effectif >=20)	Distribution écart absolu VA								
	[0,5[	39%	38%	33%	41%	25%	29%	39%	33%
	[5,10[	30%	27%	26%	27%	22%	25%	29%	27%
	[10,15[	13%	14%	18%	13%	17%	18%	17%	18%
	[15,25[	12%	13%	15%	13%	17%	18%	11%	15%
	[25,max[	6%	8%	8%	6%	20%	10%	4%	8%

Odds ratio des modèles

		Apprentissage				Apprentis en EPLE	Voie professionnelle scolaire		
Niveau de diplôme		niveau 3	niveau 4 BP	niveau 4 autres	niveau 5	tous niveaux	niveau 3	niveau 4	niveau 5
niveau de diplôme (ref= CAP)	niv 4 BP					2,89 ***			
	niv 4 bac pro					1,86 ***			
	niv 5 bts					1,53 ***			
regroupement specialités de formation (ref=2)	regroupement 1 (tx emploi faible)	0,91 .	0,98	0,74 **	0,76 ***	0,95	0,76 ***	0,8 ***	0,59 ***
	regroupement 3	1,18 ***	1,04	1,13 **	1,14 ***	1,15 *	1,13 **	1,19 ***	1,33 ***
	regroupement 4 (tx emploi fort)	1,7 ***	1,04	1,53 ***	1,45 ***	1,29 ***	1,6 ***	1,75 ***	1,83 ***
regroupement NAF (ref=2)	regroupement 1 (tx emploi faible)	0,45 ***	0,3 ***	0,36 ***	0,51 ***	0,38 ***			
	regroupement 3 (tx emploi fort)	1,5 ***	1,75 ***	1,74 ***	1,59 ***	1,86 ***			
regroupement age croissant (ref= 1)	regroupement 2	1,48 ***	0,9 .	1,18 **	0,96	1	1,72 ***	1,18 ***	0,98
	regroupement 3	1,48 ***	0,85 **	1,09 .	0,87 ***	0,96	2,06 ***	1,12 **	0,92 *
	regroupement 4	1,24 ***	0,76 ***	1,09	0,84 ***	0,79 **	2,01 ***	1,11 .	0,8 ***
mention complémentaire (ref=non)		1,7 ***		1,38 ***		1,2 .	1,92 ***	1,37 ***	
sexe (ref=homme)		0,77 ***	1,04	0,94 .	1,14 ***	0,96	0,65 ***	0,96 *	1,12 ***
PCS responsable (ref=cadres prof. intellectuelles sup.)	agriculteurs	1,14	1,33 .	1,05	1,24 *	1,39	1,43	1,35 *	1,28 *
	artisans, commerçants, chef Entre	1,22 **	1,33 **	1,14	0,97	0,86	1,4 **	1,23 ***	1,26 ***
	Professions intermédiaires	1,13 *	1,23 *	1,12 .	1,02	1,22 *	1,1	1,07	1,24 ***
	employés	1,14 *	1,21 *	1,1	0,96	1,05	1,1	1,1 *	1,24 ***
	ouvriers	1,15 **	1,37 ***	1,18 *	1,08	1,21 *	1,02	1,12 **	1,29 ***
	retraites	0,86	0,81	0,86	1,04	1,16	0,84	0,91	0,96
	sans activité professionnelle	0,97	1,1	0,9	0,88 *	0,99	0,88	0,89 **	1,05
	non réponse	1,25 ***	0,91	0,85 *	0,79 ***	0,88	1,37 ***	1	1,05
situation avant apprentissage (ref=collège)	2nd cycle gt, enseignement sup	1	0,86 **	0,8 ***	0,84 *	0,98			
	2nd cycle professionnel	1,07 *	0,87 *	0,81 ***	0,9	1			
	stage, emploi, ctt pro.	1,04	0,88	0,92	0,86 .	1,18			
	chomage	1,04	0,65 ***	0,67 ***	0,72 **	0,86			
	autres situations	1,1 *	0,84 *	0,81 ***	0,79 **	1,06			
obtention du diplôme et résultats à l'examen (ref=diplôme non obtenu)	non réponse	1,28 ***	1,07	1,75 ***	1,03	1,28 **	1,21 **	1,06	1,04
	diplôme obtenu, note = [10,12[	1,64 ***	1,84 ***	1,62 ***	1,03	1,68 ***	1,41 ***	1,41 ***	1,03
	diplôme obtenu, note = [12,14[	2,06 ***	2,1 ***	2,05 ***	1,13	2,26 ***	2,12 ***	1,65 ***	1,1
	diplôme obtenu, note = [14,20[	2,65 ***	2,15 ***	2,36 ***	0,96	2,56 ***	3,07 ***	2,06 ***	0,94
bénéficie reconnaissance travailleur handicapé (ref=non)		0,77 **	0,77	0,68 *	0,64 **	0,8			
part d'élèves en situation de handicap							0,99 ***	0,99 *	1
taux de chômage de la zone d'emploi		0,93 ***	0,94 ***	0,95 ***	0,95 ***	0,96 ***	0,9 ***	0,91 ***	0,94 ***

Significativité : *** 0,001 ** 0,01 * 0,05 . 0,1

Les variables retenues

Regroupement de spécialités de formation

Les spécialités de formation sont trop nombreuses pour être intégrées telles qu'elles dans le modèle. Elles ont été regroupées de manière automatique en fonction des taux d'insertion professionnelle observés dans InserJeunes.

Pour ce faire, nous calculons dans un premier temps les taux d'emploi par nomenclature des spécialités en 100 postes ainsi que l'effectif de sortants associé. Nous effectuons ensuite une Classification Ascendante Hiérarchique (CAH) de ces taux d'emploi pondéré par l'effectif de jeunes sortants. La méthode utilisé dans la CAH est celle de Ward qui cherche à minimiser l'inertie intra-classe et maximiser l'inertie inter-classe afin d'obtenir des classes les plus homogènes possibles. Comme InserJeunes est un système d'information où tous les calculs sont automatisés, le nombre de groupes (cluster) choisis doit être fixe et suffisamment petit pour que chaque groupe ait toujours un nombre suffisant de jeunes. En effet, cette variable étant utilisée dans les modèles multiniveaux il est important d'avoir un effectif par modalité suffisamment important.

In fine, le regroupement de spécialités de formation en 100 postes comprend 4 modalités classées du taux d'emploi le plus faible au taux d'emploi le plus élevé.

Regroupement de codes NAF (nomenclature d'activités française)

De manière similaire au regroupement de spécialités de formation, les codes NAF de l'établissement d'apprentissage sont regroupés de manière automatique. Finalement, le regroupement de codes NAF comprend 3 modalités classées du taux d'emploi le plus faible au taux d'emploi le plus élevé.

Regroupement de la variable âge en classes

Les âges étant différents selon les différents modèles réalisés par niveau de formation, les regroupements d'âge sont spécifiques à chaque modèle, mais très proches entre l'apprentissage et la voie professionnelle scolaire.

Pour le champ apprentissage, le regroupement est le suivant :

diplôme	**bornes**	**libellé**
BTS et autres diplômes de niveau 5	[min,20]	regroupement 1 (les plus jeunes)
	[21,21]	regroupement 2
	[22,23]	regroupement 3
	[24,max]	regroupement 4 (les plus âgés)
BP	[min,19]	regroupement 1 (les plus jeunes)
	[20,20]	regroupement 2
	[21,22]	regroupement 3
	[23,max]	regroupement 4 (les plus âgés)
Bac pro et autres diplômes de niveau 4	[min,19]	regroupement 1 (les plus jeunes)
	[20,20]	regroupement 2
	[21,22]	regroupement 3
	[23,max]	regroupement 4 (les plus âgés)
CAP et autres diplômes de niveau 3	[min,18]	regroupement 1 (les plus jeunes)
	[19,19]	regroupement 2
	[20,21]	regroupement 3
	[22,max]	regroupement 4 (les plus âgés)

Pour le champ apprenti en EPLE, le regroupement est le suivant :

diplôme	bornes	libellé
tous	[min,19]	regroupement 1 (les plus jeunes)
	[20,20]	regroupement 2
	[21,22]	regroupement 3
	[23,max]	regroupement 4 (les plus âgés)

Pour le champ des élèves en voie professionnelle scolaire, le regroupement est le suivant :

diplôme	bornes	libellé
BTS et autres diplômes de niveau 5	[min,20]	regroupement 1 (les plus jeunes)
	[21,21]	regroupement 2
	[22,22]	regroupement 3
	[23,max]	regroupement 4 (les plus âgés)
Bac pro et autres diplômes de niveau 4	[min,18]	regroupement 1 (les plus jeunes)
	[19,19]	regroupement 2
	[20,20]	regroupement 3
	[21,max]	regroupement 4 (les plus âgés)
CAP et autres diplômes de niveau 3	[min,17]	regroupement 1 (les plus jeunes)
	[18,18]	regroupement 2
	[19,19]	regroupement 3
	[20,max]	regroupement 4 (les plus âgés)

Mention complémentaire : On ajoute une variable indicatrice telle que si le diplôme correspond à une mention complémentaire la valeur vaut 1 et 0 sinon.

Sexe de l'élève : On ajoute une variable indicatrice telle que si le sexe correspond à une fille la valeur vaut 1 ; 0 sinon.

Profession et catégorie socioprofessionnelle

Nous prenons en compte la PCS en une position du responsable de l'élève. Nous n'effectuons pas d'imputation des non-réponses (moins de 2 % des élèves) mais créons une modalité « non réponse ».

libellé
agriculteurs
artisans, commerçants, chef d'entreprise
cadres, professions intellectuelles supérieures
professions intermédiaires
employés
ouvriers
retraités
sans activité professionnelle
non réponse

Situation avant l'apprentissage (pour les apprentis)

libellé
collège
second cycle général et technique, enseignement supérieur
second cycle professionnel
stage, emploi, contrat de professionnalisation
chômage
autres situations

Obtention du diplôme et résultats à l'examen

Obtention du diplôme	moyenne	libellé
Non réponse		Non réponse
non		diplôme non obtenu
oui	[min,12[	diplôme obtenu, note = [10,12[
	[12,14[	diplôme obtenu, note = [12,14[
	[14,max[	diplôme obtenu, note = [14,20[

Handicap

Les modèles multiniveaux prennent également en compte la spécificité de l'insertion professionnelle des jeunes sortants en situation de handicap sous la forme suivante :

- la reconnaissance de la qualité de travailleur handicapé (RQTH) pour les sortants d'apprentissage est introduite au niveau individuel.

- pour les modèles sur les jeunes sortant pour la voie professionnelle scolaire en lycée, la part d'élèves en situation de handicap a été introduite. Cette variable est issue d'une enquête recensant annuellement les élèves ayant un Projet Personnalisé de Scolarisation au sein des établissements. Elle est disponible sur le champ des élèves en année terminale de formation pour les niveaux de diplômes suivants : CAP en deuxième année, terminale professionnelle, BTS en deuxième année.

L'ajout de ces variables dans les modèles multiniveaux a été testé et a un impact positif et important sur la valeur ajoutée des CFA qui accueillent majoritairement des apprentis avec RQTH ainsi que des lycées professionnels avec une part importante d'élèves en situation de handicap.

Taux de chômage

Dans les modèles, c'est le taux de chômage annuel au niveau de la zone d'emploi qui a été retenu car les taux de chômage trimestriels ne sont pas disponibles pour les départements d'outre-mer (les DOM, excepté Mayotte, font partie du champ d'InserJeunes). La comparaison, au niveau France métropolitaine, de l'ajout du taux de chômage annuel par rapport au trimestriel a montré un impact très modéré.