Les appariements de données de la statistique publique : des analyses enrichies, un cadre juridique protecteur

Publié le 1er septembre 2023 sur le blog de l'Insee
Temps de lecture : 10 minutes

Françoise Dupont, Insee.

L'Insee, et plus largement le service statistique public, collecte des informations d'origines diverses, notamment par des enquêtes sur échantillons ou par la réutilisation à des fins statistiques de données administratives. Ces données peuvent être utilisées seules, ou combinées à d'autres sources, au niveau individuel, pour fournir une information plus riche. C'est ce que l'on appelle « faire des appariements ». Cette pratique répond à des besoins très divers ; elle est pratiquée de longue date, en France comme dans d'autres pays, et strictement encadrée d'un point de vue juridique.

Qu'est-ce qu'un appariement de données ?

Apparier ou croiser des données relatives aux individus consiste à rassembler pour une même personne ou entreprise[1] des données qui la concernent et qui sont issues de différentes sources.

1 On peut également apparier des données portant sur d'autres objets d'études comme les logements, les entreprises, des régions, des pays, etc. Dans ce cas on rassemble les informations sur un logement, une entreprise… donné(e) issues de plusieurs sources.

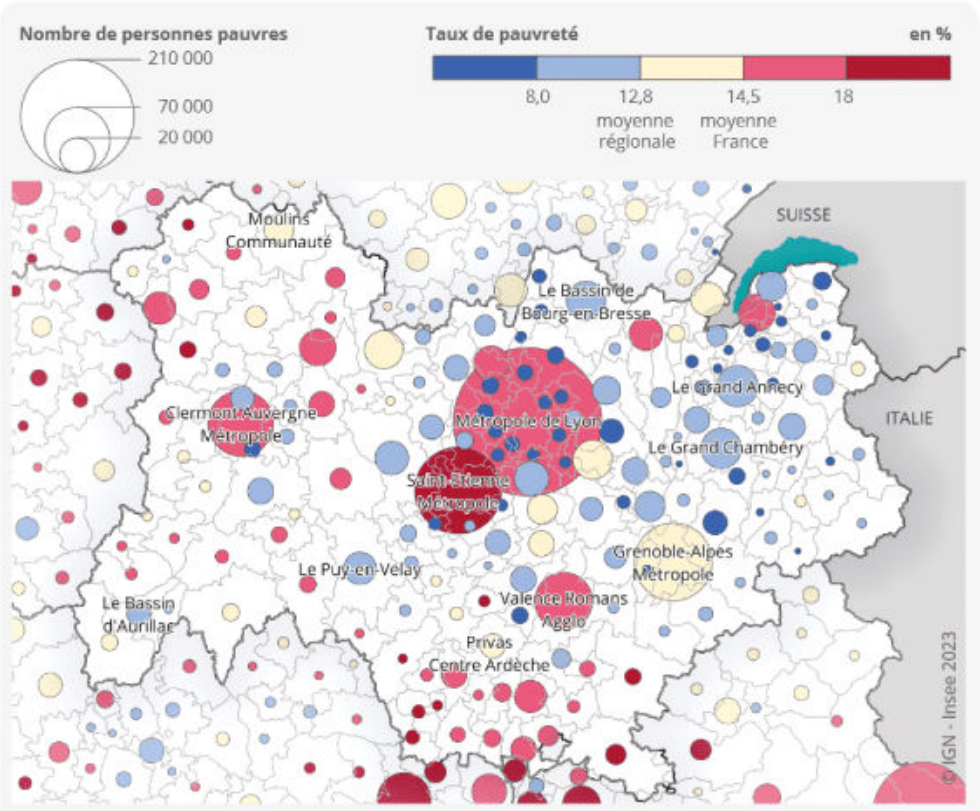
Pour quelles raisons utilise-t-on des appariements de données ?

On utilise des appariements de données parce qu'une source seule n'est pas toujours suffisante et ne couvre qu'une partie de ce que l'on veut analyser. Prenons quelques exemples :

Pour construire des statistiques complètes sur les revenus des ménages, on utilise des données issues des déclarations de revenus à l'impôt sur le revenu, mais on a également besoin des informations tirées de fichiers d'allocataires de prestations familiales et sociales (CAF, MSA, Cnav) qui permettent d'ajouter les prestations versées à l'ensemble des ménages. L'Insee a ainsi mis en place le dispositif Filosofi (Fichier localisé social et fiscal), qui permet d'avoir une vue plus complète des revenus des ménages en rapprochant les deux sources de données pour chaque individu. On peut alors analyser les disparités de revenus entre territoires et au niveau de territoires particuliers, par exemple une région, un département ou même les quartiers d'une ville *(figure 1)*.

Figure 1 – Nombre de personnes pauvres et taux de pauvreté par intercommunalité en Auvergne-Rhône-Alpes en 2019

Figure 1 - Nombre de personnes pauvres et taux de pauvreté par intercommunalité en Auvergne-Rhône-Alpes en 2019



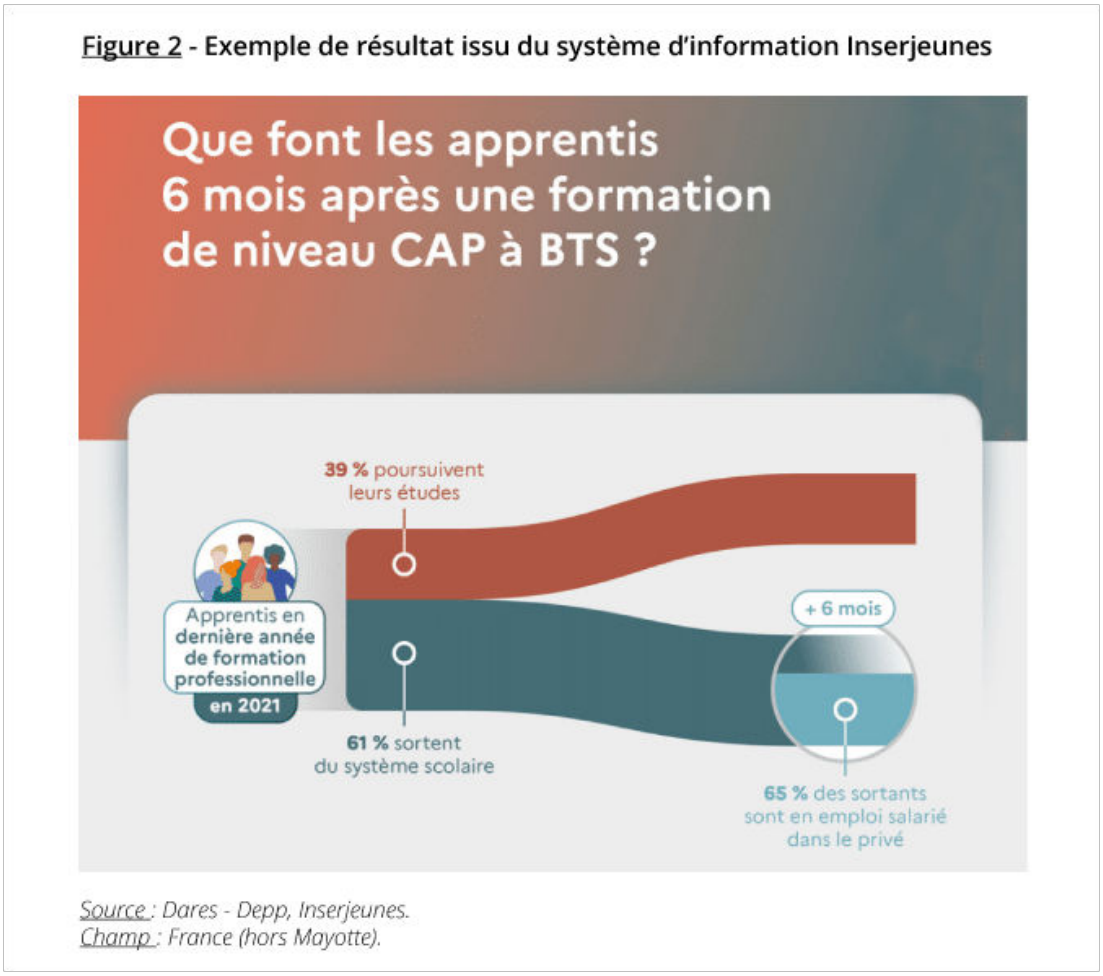
Lecture : en 2019, la Métropole de Lyon compte 213 800 personnes pauvres vivant dans des ménages fiscaux ordinaires. Son taux de pauvreté s'élève à 16,2 %. Le taux de pauvreté régional est de 12,8 %.
Source : Insee-DGFIP-Cnaf-Cnav-CCMSA, Fichier localisé social et fiscal (Filosofi) 2019.

Pour bâtir des statistiques exhaustives sur les montants de retraites, on a besoin de rassembler les données des différents régimes : Cnav pour les salariés, MSA pour les agriculteurs y compris salariés, SSI pour pour les artisans et commerçants, régimes spéciaux. C'est ce que fait l'échantillon interrégimes de retraités (EIR) construit par la Drees, le service statistique de la santé et des solidarités. Il apparie, pour un échantillon de retraités, des données sur leurs montants de retraite dans les différents régimes qui les concernent, afin de reconstituer leurs montants de retraite totale. Il offre ainsi la possibilité de mener des analyses approfondies sur l'évolution et les disparités de montants de retraite perçus, en particulier selon la génération ou le sexe [Drees, 2023].

Lorsque l'on veut évaluer l'impact d'une aide sociale, ou d'une aide à destination des entreprises, on apparie les données dont on dispose sur les bénéficiaires de l'aide avec un fichier qui décrit leur situation avant et après l'aide (par exemple l'emploi ou la réussite dans l'enseignement supérieur pour une personne, les résultats financiers pour une entreprise). On compare avec une situation de référence de personnes ou d'entreprises n'ayant pas bénéficié de l'aide. On peut en déduire si l'aide a été bénéfique globalement, et dans quel type de situation elle a porté ses fruits.

Lorsque l'on veut reconstituer des parcours des personnes. Par exemple, pour décrire l'insertion professionnelle des jeunes, on ajoute des informations sur l'emploi à des informations décrivant le parcours scolaire des jeunes diplômés des centres de formation d'apprentis (CFA) et des lycées professionnels. Le système d'information Inserjeunes, porté par la Depp et la Dares (respectivement les services statistiques chargés de l'éducation et de l'emploi), rapproche ainsi un extrait de la base de données administratives constituée pour la gestion de la scolarité de ces élèves et un extrait des données sur l'emploi issues des déclarations sociales nominatives qui sont remplies par les employeurs *(figure 2)*. Ce travail d'appariement permet d'analyser les filières d'apprentissage qui facilitent l'accès au marché du travail (et donc de mieux guider les jeunes), ou encore de connaître avec précision le nombre de jeunes qui abandonnent leur apprentissage en cours de route.

Figure 2 – Exemple de résultat issu du système d'information Inserjeunes



Source : Dares – Depp, Inserjeunes.
Champ : France (hors Mayotte).

Autre exemple, pour mieux comprendre les difficultés des personnes bénéficiaires des minima sociaux et l'évolution de leur situation, on va apparier les données de revenus sur plusieurs années, avec les données d'emploi sur plusieurs années et obtenir ainsi des évolutions individualisées qu'on peut ensuite analyser. La reconstitution des trajectoires individuelles des bénéficiaires de minima sociaux a ainsi permis de mieux comprendre l'entrée dans la pauvreté et d'évaluer les dispositifs de protection sociale.

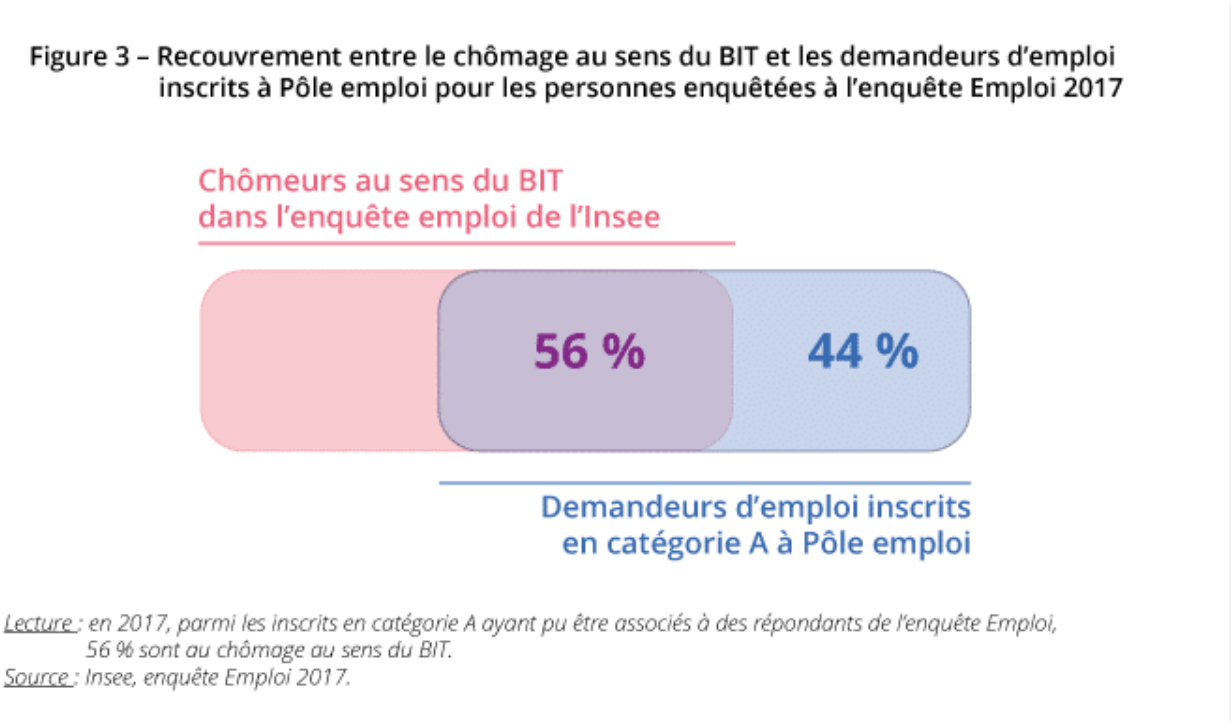
Lorsque l'on veut alléger les questionnaires d'enquête. Les statisticiens publics cherchent à éviter de demander à un ménage (ou à une entreprise) une information qu'il (elle) a déjà transmise à une administration, en particulier si elle est complexe et longue à reconstituer. Par exemple, dans l'enquête logement, on concentre les questions sur le thème de l'enquête qui est déjà riche et on ajoute par appariement, une fois l'enquête terminée, le revenu issu de bases de données sur les revenus.

Autre exemple, en 2009, les enquêtes annuelles auprès des entreprises ont connu un allègement très significatif grâce à une plus grande utilisation des données fiscales et sociales portant sur les entreprises. Ces dernières permettent de connaître pour chaque entreprise, sans lui redemander l'information, le chiffre d'affaires, l'excédent brut d'exploitation, la valeur ajoutée, le résultat comptable,

les immobilisations, la marge commerciale, l'emploi total, l'emploi salarié, l'emploi non salarié, etc. Le dispositif Esane (Élaboration des statistiques snnuelles d'entreprise) combine ainsi des données administratives (obtenues à partir des déclarations annuelles de bénéfices que font les entreprises à l'administration fiscale, et à partir des déclarations sociales qui fournissent des informations sur l'emploi) et des informations obtenues à partir d'un échantillon d'entreprises enquêtées annuellement par un questionnaire spécifique.

Pour mieux comprendre certains phénomènes comme le chômage : en appariant les données sur les demandeurs d'emploi inscrits à Pôle emploi avec celles de l'enquête Emploi réalisée par l'Insee, on a pu décrire et quantifier toutes les situations qui expliquent pourquoi le chômage au sens du Bureau international du Travail, mesuré par l'enquête Emploi, évolue différemment du nombre de demandeurs d'emploi inscrits à Pôle emploi *(figure 3)*.

Figure 3 – Recouvrement entre le chômage au sens du BIT et les demandeurs d'emploi inscrits à Pôle emploi pour les personnes enquêtées à l'enquête Emploi 2017



Lecture : en 2017, parmi les inscrits en catégorie A ayant pu être associés à des répondants de l'enquête Emploi, 56 % sont au chômage au sens du BIT.
Source : Insee, enquête Emploi 2017.

Pour toutes ces raisons, les techniques d'appariement permettent d'élaborer des statistiques plus riches et / ou limitant le fardeau de réponse, favorisant une meilleure connaissance de la société dans toute sa complexité et sa diversité, pour les pouvoirs publics, la recherche et la société civile.

Depuis quand l'Insee et les statisticiens publics utilisent des appariements de données ?

L'Insee et les Services statistiques ministériels qui constituent le service statistique public ont une longue expérience d'appariements de données sur les individus, qui s'est construite au fil du temps.

Dès la fin des années 1950, l'enquête Revenus fiscaux rapprochait des données du recensement de la population et des données administratives d'origine fiscale pour un échantillon de personnes, afin d'établir des statistiques sur les revenus des ménages et les inégalités.

L'échantillon démographique permanent (EDP) mis en place en 1967 apparie des données issues des recensements de la population et de l'état civil. Il s'est peu à peu enrichi par appariement avec de nouvelles sources. Il permet par exemple d'étudier la mobilité intergénérationnelle en termes de revenus [Abbas et Sicsic, 2022], de mettre en évidence le rôle protecteur du couple lors de la perte d'emploi [Fabre et Lacour, 2021] ou de mesurer les changements de résidence au moment du départ en retraite [Abbas *et alii*, 2022]. Récemment, il a été apparié par le service statistique ministériel chargé de la santé et des solidarités (Drees) avec des données issues du *Système national des données de santé* pour construire l'EDP santé.
Au cours des années 1970, différents panels sont mis en place pour étudier des parcours dans le temps :

- **Les parcours scolaires dès 1973,** avec des panels d'élèves basés sur les données administratives issues de la gestion du système scolaire pour un échantillon d'élèves que l'on suit dans le temps. Ils ont permis d'abord d'observer les entrants en sixième, puis les élèves entrant en cours préparatoire à partir de 1978, puis ceux entrant en petite section à partir de 2021. Ils se sont enrichis depuis les années 1990 avec des évaluations des acquis et des enquêtes auprès des élèves ou des familles [Depp, 2017].

- **Les carrières salariales dès 1976**. Un panel permet de suivre dans le temps, pour un large échantillon de personnes, les emplois qu'elles occupent et les salaires perçus. D'abord limité au secteur privé, le panel a été ensuite étendu dans les années 2000 au secteur public (panel « tous salariés »), puis aux non-salariés (panel « tous actifs »).

Les appariements de données se sont ensuite développés peu à peu au fil du temps en fonction des besoins de connaissance, favorisés également par des capacités informatiques de traitement de grosses bases de données en extension et par un accès croissant des services statistiques, sur des bases légales, à des bases de données administratives plus nombreuses.

Pour les données portant sur les entreprises, l'existence d'un identifiant unique et partagé au niveau de l'entreprise (Siren) ou de l'établissement (Siret), géré par l'Insee dans le répertoire Sirene, permet depuis 1970 de rapprocher facilement les différentes informations disponibles pour une entreprise[2] pour produire des statistiques.

Dans une étude parue récemment en 2021, les données de chiffres d'affaires issues des déclarations mensuelles de TVA ont ainsi été appariées avec des données de la source Esane (Élaboration des statistiques annuelles d'entreprise) afin de caractériser les 600 000 entreprises de l'étude, d'analyser les évolutions d'activité pendant la crise sanitaire et de dégager quatre profils type d'impact de cette crise.

2 À noter, les données relatives à des entreprises individuelles sont considérées comme des données personnelles, dans le cas où elles sont rattachables à une personne physique identifiable

Comment réalise-t-on en pratique un appariement ?

Pour apparier deux fichiers de données, il faut se baser sur des variables communes présentes dans les deux fichiers sous la même forme ou qui peuvent être ramenées à la même forme pour être comparées. Pour des appariements concernant des personnes, il peut s'agir de l'identifiant réservé aux appariements par le service statistique public comme le Code Statistique Non Signifiant, ou encore des noms et prénoms, des prénoms et adresses, ou dans des cas plus rares et très encadrés et prévus par un décret en Conseil d'État du NIR (Numéro d'inscription au répertoire au RNIPP). Pour des appariements concernant des entreprises, on utilise l'identifiant unique et partagé au niveau de l'entreprise (Siren) ou de l'établissement (Siret).

Quel cadre juridique pour les appariements de données personnelles à des fins statistiques ?

Le cadre juridique général des appariements de données personnelles à des fins statistiques est constitué des deux lois qui encadrent toute la collecte de données et la production des statistiques du service statistique public :

- Tout d'abord la loi de 1951 sur l'obligation, la coordination et le secret en matière de statistique, qui est une sorte de « loi statistique » qui encadre l'ensemble des travaux statistiques et les autorise.

- La loi relative à l'informatique, aux fichiers et aux libertés de 1978, qui encadre le traitement informatique de ces données. Elle a été modifiée au fil du temps et en particulier pour intégrer certaines dispositions de la loi pour une République numérique de 2016 (comme la mise en place du Code statistique non signifiant), puis lors de la mise en place en 2018 au niveau européen du Règlement général sur la protection des données (RGPD), qui encadre l'utilisation des données personnelles.

Les appariements à des fins de recherche scientifique ou historique sont également prévus et encadrés par la loi, qui prévoit des dispositions spécifiques, garantissant à la fois la possibilité de réaliser de tels appariement et la protection des données concernées.

Quelle information du public ?

Parallèlement aux évolutions de ce cadre juridique, les statisticiens, en France comme à l'étranger, ont progressivement formalisé et enrichi leur réflexion sur les aspects déontologiques de ces traitements et sur la nécessité de disposer d'un « mandat social » en plus du « mandat juridique ».
C'est ainsi que les appariements de données qui sont utilisés pour construire les statistiques sont listés dans les programmes de travail qui sont présentés chaque année au Conseil national de l'information statistique, instance de concertation entre les producteurs et les utilisateurs de la statistique publique.

Que font les autres pays ?

La plupart des instituts de statistiques utilisent des appariements de données. Ils sont préconisés au niveau européen, tout comme l'emploi de sources administratives, pour limiter les coûts et réduire la

charge des personnes ayant à répondre à une enquête. Ainsi deux des principes du Code des bonnes pratiques de la statistique européenne évoquent ces objectifs : le Principe 9 de **charge non excessive pour les déclarants** « Afin d'éviter la multiplication des demandes de données, les sources administratives ou autres sont mobilisées autant que possible » et le Principe 10 de **rapport coût efficacité** « les ressources sont utilisées de façon efficiente ; tout est mis en œuvre pour améliorer l'exploitation statistique des sources de données administratives ou autres et pour limiter le recours à des enquêtes directes ».

Chez nos collègues canadiens, les appariements font l'objet d'un texte juridique spécifique depuis 1986. Ils font depuis peu l'objet d'un examen de nécessité et de proportionnalité, conformément au cadre mis en place en 2019. Le principe de nécessité vise à s'assurer qu'il y a un bénéfice pour la société canadienne associée au traitement de données. Pour le principe de proportionnalité, les experts de Statistique Canada déterminent comment recueillir uniquement les données nécessaires, en prenant en compte la nature plus ou moins sensible des données et en vérifiant qu'il n'y a pas d'alternative moins gourmande en données pour atteindre le même objectif.

Pour en savoir plus

- Dupont F., 2023, « Quelles sources utilise l'Insee pour construire ses statistiques », *Blog de l'Insee*, mai
- D'Alessandro C., Dupont F., Guillaumat-Tailliet F., 2023, « Appariements de données individuelles : vers une meilleure qualité et plus de transparence », Cnis – *Chronique*s n° 32, avril
- Cnis, 2022, « Appariements de données individuelles : entre richesse de l'information statistique et respect de la vie privée », *Rencontres du Cnis* du 28 janvier 2022

Exemples d'appariements :

- **Pôle emploi et chômeurs :**
  Coder Y., Hamman S. (Pôle emploi) ; Dixte C. (Dares) ; Hameau A., Larrieu S., Marrakchi A., Montaut A. (Insee), 2019, « Les chômeurs au sens du BIT et les demandeurs d'emploi inscrits à Pôle emploi : une divergence de mesure du chômage aux causes multiples », juillet
- **Formation et emploi (inserjeunes) :**
  Antoine R., Collin C., Marchal N. (DEPP), Fauchon A. (Dares), 2021 « Insertion professionnelle des apprentis du niveau CAP à BTS : 6 mois après leur sortie du système éducatif en 2020, 61 % sont en emploi salarié en janvier 2021 », DEPP – *Note d'information* n° 21.43, décembre
- **Les trajectoires passées des bénéficiaires de minima sociaux :**
  Cabannes P-Y., Chevalier M., 2022, « Minima sociaux et prestations sociales – Ménages aux revenus modestes et redistribution – Édition 2022 », *Panoramas de la DRESS* – fiche 21, Drees , septembre
- **Les trajectoires mensuelles d'activité d'entreprises sur 2019-2020 :**
  Giorgi J., Scott S. (Insee), 2021, « Pandémie de Covid-19 et pertes d'activité : évaluation de l'impact de la crise sur les trajectoires des entreprises françaises en 2020 », Les entreprises en France – Édition 2021, *Insee Références*, décembre