PRÉSENTATION DU NUMÉRO

Relancé en 2018 dans une nouvelle formule, le *Courrier des statistiques* atteint sa troisième année d'existence et bientôt 50 articles. Nous avons pu explorer des sujets variés, méthodes, outils, mais aussi des questions institutionnelles ou juridiques posées par la statistique publique, en veillant à rester ouverts sur l'extérieur, en France ou à l'étranger, afin de se comparer et de nourrir nos réflexions.

La revue ne peut faire l'impasse sur un changement majeur de ces dernières années : si les statisticiens continuent à organiser la collecte d'informations par enquête, et à innover en la matière[1], ils doivent aussi, de plus en plus, tirer parti d'un monde où des données existent déjà, qu'ils n'ont pas construites. On pourra objecter que cela a toujours existé avec les sources administratives, mais celles-ci évoluent, s'enrichissent, comme on l'a vu dans le numéro N1, à travers la DSN[2] notamment. Plus généralement, les données externes font désormais toujours partie du paysage.

L'interrogation sur les gisements de données existants, leur mode d'obtention, leur degré d'élaboration, leur champ, leur temporalité, est systématique. Il s'agit là d'une préoccupation centrale dans le présent numéro, sixième de la nouvelle série, qui commence en présentant justement quatre sources de données, tout à fait essentielles pour des usages statistiques.

À tout seigneur tout honneur, c'est l'enquête Emploi qui ouvre le bal : enquête phare de la statistique publique en France, autour de laquelle gravitent d'autres opérations statistiques, l'enquête Emploi reste année après année une source inépuisable pour les études socio-économiques. Pour autant, celle-ci n'est pas immuable, elle s'adapte à un monde qui change. **François Guillaumat-Tailliet et Chloé Tavan** présentent les grandes lignes de la refonte de 2021 et ses motivations : l'exigence d'harmonisation au niveau européen[3], mais aussi la volonté de développer la possibilité pour les ménages d'utiliser internet pour répondre aux enquêtes de l'Insee. Cette somme de changements significatifs a nécessité un long travail de préparation et d'expérimentation, ainsi qu'une vaste opération préalable pour estimer les ruptures de séries.

Il fallait bien un jour ouvrir les colonnes de la revue aux promoteurs de Fidéli, souvent cité dans les numéros précédents : le fichier démographique sur les logements et les individus n'est ni le résultat d'une enquête, ni une source administrative *stricto sensu*. Comme l'expliquent **Pierre Lamarche et Stéfan Lollivier**, Fidéli est une pure construction des statisticiens pour leurs propres besoins, un travail de mise en cohérence, et d'enrichissement de sources administratives, notamment fiscales. La cohérence, l'exhaustivité et la variété d'informations disponibles sont essentielles pour son insertion dans le système d'information du Service statistique public. À partir de plusieurs sources « brutes », le dispositif élabore une liste unique de logements d'habitation et une liste unique d'individus, localisés dans leur logement principal, tout en regroupant les informations socio-démographiques les concernant. Fidéli permet de réaliser des études spécifiques et d'échantillonner des enquêtes, mais aussi de compléter, par appariement, des données d'enquêtes avec des données socio-démographiques finement localisées, ce qui démultiplie son potentiel pour l'analyse sociale.

1. Voir le *numéro N3*.
2. Déclaration sociale nominative, voir le *numéro N1*.
3. Dont nous avions annoncé le cadre dans le *numéro N3* et qui fait écho à d'autres refontes (voir *numéro N2*).

L'échantillon démographique permanent (EDP) apporte une corde supplémentaire à notre arc : la profondeur temporelle, la possibilité de travailler sur des cohortes, de mieux comprendre des évolutions dans le temps et au fil des générations. C'est une source déjà ancienne, qui retrace pas moins de 3,7 millions de trajectoires individuelles, dont 200 000 depuis plus de 50 ans. **Isabelle Robert-Bobée et Natacha Gualbert** en décrivent les fonctionnalités actuelles − car son contenu et son périmètre n'ont cessé de s'étoffer au fil des années − et les principales innovations − car il a fallu également que l'EDP s'adapte aux évolutions de son environnement. Il s'est ainsi enrichi récemment de données socio-fiscales, comme Fidéli. La compilation de différentes sources fait l'originalité de l'EDP : si elle rend plus complexe son exploitation pour les études, elle offre en retour des possibilités d'analyse des trajectoires de plus en plus diversifiées.

Continuant l'exploration du vaste univers des sources administratives, **Christian Sureau et Richard Merlen** abordent le répertoire général des carrières unique (RGCU) mis au point par la Cnav (Caisse nationale d'assurance vieillesse). Cette gigantesque base de données doit, à terme, permettre aux organismes de retraite, aux entreprises, à leurs salariés et aux retraités, de partager une information respectant les mêmes concepts, sur les différentes dimensions des périodes qui composent une carrière professionnelle (activité salariée, période de chômage, etc.). La qualité de ce répertoire est assurée, entre autres, par un mécanisme sophistiqué de contrôle des données, à plusieurs niveaux. Au départ outil au service des usagers et de l'efficience administrative, voici une base de données qui est promise à un bel avenir pour des usages statistiques, de par sa richesse sans équivalent. Elle incorpore, comme l'EDP, une dimension temporelle : on remonte même bien plus loin, jusqu'aux années trente. Contrairement aux trois autres sources, le RGCU n'est pas encore disponible, et il faudra patienter encore quelques mois jusqu'en 2022 pour qu'une première version soit mise à disposition des chercheurs *via* le CASD.

Pour répondre aux besoins d'évaluation de la puissance publique, les services statistiques ministériels ont à leur disposition des sources de données, mais les relier entre elles n'a rien d'immédiat dès lors qu'on ne dispose pas d'identifiant commun. Cette activité d'« appariement » de fichiers (*record linkage* en anglais) a suscité depuis les années 80-90 une vaste littérature académique, notamment au Canada, aux Pays-Bas, en Australie, aux États-Unis, ou en Italie. Avec la multiplication des sources administratives disponibles, l'intérêt pour ces méthodes et leur mise en application s'accroît au sein de la statistique publique française.

Les travaux menés par la Depp s'inscrivent dans cette mouvance. InserJeunes, dispositif sur l'insertion professionnelle des jeunes, est ainsi basé sur l'appariement de sources administratives exhaustives, relatives à la scolarité des élèves et des apprentis, à la réussite aux examens, aux contrats d'apprentissage et aux contrats salariés, et de la déclaration sociale nominative (DSN). L'article de **Loïc Midy** explicite les étapes nécessaires pour mener à bien l'appariement de ces fichiers sur identifiant indirect : normalisation, indexation, calcul de similarités, classification des paires et évaluation/validation. Il met en évidence au passage la complexité de l'opération et les limites des approches « naïves » de ce sujet. L'auteur s'interroge également sur la manière d'outiller le processus, en effectuant un tour d'horizon des outils d'appariement *open source* existants.

S'intéresser aux données administratives, c'est aussi s'interroger sur la qualité de ces données, ne pas considérer celle-ci comme un acquis. De ce point de vue, la Belgique a clairement un temps d'avance, avec un travail de fond mené sur les anomalies dans les données et la manière d'y remédier en remontant à la source même de l'information : l'intérêt

de la démarche a même été acté par un « arrêté royal » qui l'impose aux administrations. **Isabelle Boydens, Gani Hamiti et Rudy Van Eeckhout** nous présentent ainsi un prototype original, appelé ATMS (*Anomalies & Transactions Management System*). Il permet un suivi des anomalies et des traitements, en support à la méthode dite du *back tracking* : dans une approche préventive de la qualité des données, la méthode est destinée à améliorer structurellement la qualité à la source, et fait également le lien avec des approches curatives, plus classiques (*data quality tools*).

Enfin, le dernier article de ce numéro nous emmène dans une direction beaucoup plus institutionnelle, avec le Conseil national de l'information statistique (Cnis). Il fait écho aux articles précédemment publiés, qui portaient sur deux autres acteurs de la gouvernance statistique en France, que sont l'Autorité de la Statistique Publique d'une part, le Comité du Label d'autre part[4]. **Isabelle Anxionnaz et Françoise Maurel** nous décrivent ici les principes de fonctionnement du Cnis et nous en dévoilent les arcanes. Elles nous rappellent notamment son rôle crucial dans l'organisation de la concertation entre producteurs et utilisateurs de statistiques publiques. Elles démontrent qu'institution n'est pas synonyme d'organisation figée : les rencontres et les groupes de travail du Cnis produisent, en toute transparence, une vision partagée sur les besoins de statistiques et sur la pertinence des productions. Et au-delà, les recommandations sont mises en application dans les programmes de la Statistique publique. Le Cnis contribue enfin, de par sa vigilance aux évolutions des usages, à l'adaptation en continu du Service statistique public dans ses méthodes de production, à leur rationalisation et à leur articulation avec les sources existantes.

Odile Rascol
Rédactrice en chef, Insee

4. Voir le *numéro N5*.