PRÉSENTATION DU NUMÉRO

Ces dernières années, l'univers de la donnée s'est considérablement transformé, ouvrant la voie à nombre d'innovations, organisationnelles ou scientifiques. Le Courrier des statistiques s'attache régulièrement à éclairer ces évolutions, voire à éclairer comment naissent les innovations. Vous souvenez-vous de l'article « Naissance d'une innovation en production statistique » de Jean-Marc Béguin dans le numéro N3 ?

Dans ce numéro N13, nous vous emmenons au-delà du système statistique public. En effet, la science des données au service de missions d'intérêt public ne s'arrête pas à ses frontières. Récemment, dans un contexte de multiplication des données et de démocratisation des méthodes pour les traiter, plusieurs inspections générales se sont dotées d'un pôle de science des données. Pour ouvrir ce numéro, Quentin Bolliet, Aymeric Floyrac, Sophie Maillard et Agathe Rosenzweig présentent le pôle science des données de l'inspection générale des finances (IGF), créé en 2019. Puis, Juliette Berthe présente le pôle *data* de l'inspection générale des affaires sociales (Igas), créé en 2023. Au-delà de la spécificité de chacun de ces pôles, ils ont en commun le cadre particulier dans lequel ils interviennent : celui de missions visant à répondre à des questionnements de politique publique souvent très ciblés et dans des délais parfois fortement contraints. À la différence du système statistique public, ils réalisent des travaux sur mesure et de court terme, même s'ils s'attachent, autant que possible, à capitaliser au fil des missions les investissements réalisés.

Le pôle science des données de l'IGF compte aujourd'hui une dizaine d'agents. Il intervient en appui aux inspecteurs des finances, voire, pour certaines missions, seul. Il s'agit souvent de missions d'évaluation de politique publique, comportant une première phase de diagnostic et une deuxième phase de simulation de réforme. Le pôle s'appuie largement sur les données du système statistique public (mais pas seulement) et sur les infrastructures développées par l'écosystème (comme le Centre d'accès sécurisé aux données [N3] ou les plateformes de *data science* SSPCloud [N7] et Nubonyxia). La force du pôle est sa polyvalence, sa capacité à mobiliser des sources et des méthodes quantitatives très variées pour éclairer le plus finement possible les questionnements très précis des missions. Son rôle est aussi d'accompagner les résultats des analyses avec la pédagogie nécessaire, afin que les inspecteurs des finances et les décideurs publics puissent en percevoir au mieux les enseignements et les limites. L'exemple que les auteurs présentent sur l'assurabilité des collectivités territoriales illustre la grande technicité des travaux menés par le pôle au service d'investigations à la fois très concrètes et d'une grande actualité.

Le pôle *data* de l'Igas, plus récent, a appuyé une vingtaine de missions sur des thématiques très variées. L'Igas opère dans trois domaines stratégiques pour la vie des citoyens : la santé, le travail et les solidarités. Afin d'éclairer les questionnements de politique publique les concernant, le pôle peut mobiliser de nombreuses données très structurées dans les domaines de la santé et du travail, à commencer par le système national des données de santé (SNDS) et les sources basées sur la déclaration sociale nominative (DSN) [N1]. Dans le domaine des solidarités, où interviennent de multiples acteurs, les données sont beaucoup moins centralisées. Pour certains sujets très précis, le pôle est parfois amené à exploiter directement les données des systèmes de gestion des acteurs locaux, avec toutes les difficultés que pose le recours à des sources non dédiées en premier lieu à un usage statistique. Parfois, il n'existe aucune base de données disponible pour répondre à la mission. Le pôle peut alors être amené à créer ses propres bases, en recourant à des

techniques telles que le *web scraping*. À travers de nombreuses illustrations, l'autrice met en avant la diversité des situations rencontrées et, à travers elles, l'enjeu de l'existence de bases de données structurées et standardisées.

Dans la lignée des articles consacrés aux grands outils du système statistique public, le voyage se poursuit avec la découverte du code officiel géographique (COG). Souvent confondu avec le code postal, il est pourtant présent dans nos vies depuis plus longtemps, puisqu'il se niche dans notre numéro de sécurité sociale. Le fait qu'il ait été consacré jeu de données de référence par la loi pour une République numérique illustre d'ailleurs sa discrète essentialité. Le COG, c'est un ensemble de listes de territoires, des communes jusqu'aux pays, avec un code qui permet d'identifier de manière unique chacun d'eux « à une date donnée ». Car les territoires peuvent évoluer : des communes se créent, fusionnent, disparaissent, etc. Joachim Clé, Frédéric Minodier, Violaine Simon et Pierre Vernédal racontent l'histoire de ce répertoire, qui date d'avant la création de l'Insee, et à travers elle celle de la France et de ses découpages territoriaux. Ils mettent en avant les usages importants qui en sont faits par les administrations et expliquent le processus minutieux et multipartenarial qui permet de le mettre à jour chaque année et d'en assurer une diffusion efficace et moderne.

Enfin, ce numéro consacre un dossier aux explorations menées par l'Insee de données détenues par des opérateurs privés. C'est en 2010 que l'Insee utilise pour la première fois ce type de source : il s'agit alors de « données de caisse », c'est-à-dire d'informations recueillies par les enseignes du commerce de détail, au moment où les clients passent à la caisse, sur les produits achetés et les prix payés. Marie Leclair retraçait dans le numéro N3 la chronologie de ce projet qui a finalement conduit, en janvier 2020, à rénover profondément la méthode d'élaboration de l'indice des prix à la consommation. Dans le numéro N12, vous avez pu découvrir les travaux menés par l'Insee à partir de données de comptes bancaires.

Dans le premier article de ce dossier, Romain Lesur dresse un panorama des explorations de données d'opérateurs privés menées par l'Insee au-delà des données de caisse : téléphonie mobile, plateformes d'hébergement de courte durée, relevés de comptes bancaires, mais aussi transactions par carte bancaire. Toutes ces sources présentent un fort potentiel pour compléter les sources traditionnelles du système statistique public, grâce à leur fine granularité temporelle et spatiale. En revanche, elles posent des difficultés pour un usage à des fins d'élaboration de statistiques publiques. Le fait qu'il s'agisse de données massives n'est plus aujourd'hui la question première : l'Insee maîtrise les méthodes de traitement de telles données, dites méthodes de *data science*, et dispose des infrastructures adaptées. Les interrogations actuelles portent davantage sur la manière d'organiser un partenariat durable entre l'institut et les opérateurs privés, sur le cadre juridique dans lequel ce partenariat peut s'inscrire et sur le processus à imaginer pour rendre les données exploitables, tout en veillant à respecter strictement leur confidentialité. L'Europe s'est emparée de ces questions : l'auteur présente les grandes évolutions législatives et les projets en cours à ce niveau.

Dans le deuxième article, Marie-Pierre Joubert présente les travaux menés à partir des données de téléphonie mobile, dont les premières explorations datent de 2016. Le premier défi posé au statisticien face aux données d'opérateurs privés est de comprendre le processus par lequel elles sont recueillies, processus dont la finalité n'est pas statistique.

Grâce à plusieurs partenariats menés avec des opérateurs (Orange, mais aussi Bouygues et SFR pendant la crise sanitaire), l'Insee a pu mieux comprendre les traces numériques engendrées par les connexions aux antennes relais. Les données de téléphonie mobile se sont révélées précieuses pour éclairer les déplacements de population lors des épisodes de confinement et donner ainsi des éléments utiles aux décideurs pour cibler au mieux les besoins en services publics. Plus généralement, ces données affinent la vision des dynamiques de population, en contribuant par exemple à éclairer des mécanismes de ségrégation sociospatiale ou encore à mieux saisir les liens entre les territoires. Néanmoins, de nombreuses difficultés se posent pour gérer des problèmes d'incertitude spatiale et temporelle ou parer aux défauts de couverture ou de représentativité.

Dans le troisième et dernier article de ce dossier, Mathieu Boittelle, Émilie Cupillard, Alain Jacquot, Marie-Pierre Joubert et Florian Le Goff exposent les travaux menés à partir des données de transactions par carte bancaire CB. La plupart des transactions bancaires passent en effet par un réseau qui intermédie ces échanges entre les banques de l'acheteur et du commerçant. Le groupement d'intérêt économique Cartes Bancaires CB pilote le schéma de paiement domestique français CB, qui est le principal schéma utilisé en France devant les schémas internationaux comme Visa ou Mastercard. Depuis le printemps 2020, le groupement transmet régulièrement des données agrégées de flux de paiement CB à l'Insee. Ces dernières contribuent notamment à réaliser une estimation avancée du volume des ventes dans le commerce de détail. Des travaux de recherche menés dans le cadre de la chaire Finance digitale montrent que les données CB peuvent être précieuses pour compléter les analyses sur les connexions entre commerces et territoires. À l'instar des autres données détenues par des opérateurs privés et non destinées à des fins statistiques, elles demandent un fort investissement méthodologique pour être comprises et utilisées en tenant compte de leurs limites. Elles ne peuvent se substituer aux sources traditionnelles du système statistique public, mais apportent de nouvelles connaissances inaccessibles à partir de ces dernières.

Bonne lecture !

Emmanuelle Nauze-Fichet
Rédactrice en chef, Insee