Juin 2025

Courrier des statistiques

13

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE

Rédaction en chef
Emmanuelle Nauze-Fichet

Contribution
Insee : Mathieu Boittelle, Joachim Clé, Émilie Cupillard, Alain Jacquot, Marie-Pierre Joubert, Florian Le Goff, Romain Lesur, Frédéric Minodier, Violaine Simon, Pierre Vernédal
Igas : Juliette Berthe
IGF : Quentin Bolliet, Aymeric Floyrac, Sophie Maillard, Agathe Rosenzweig

Directeur de la publication
Jean-Luc Tavernier

Directeur de la collection
Pascal Rivière

Rédaction
Solenn Ily, Marine Le Roux, Emmanuelle Nauze-Fichet, Pascal Rivière

Composition
Agence Efil
90 boulevard Heurteloup
37 000 Tours
02 47 47 03 20
www.efil.fr

Photo de couverture
Getty Images

Éditeur
Institut national de la statistique et des études économiques
88, avenue Verdier
92541 MONTROUGE CEDEX

www.insee.fr

Courrier des statistiques N13

SOMMAIRE

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

L'inspection générale des finances et la science des données

Quelles méthodes pour quels usages ?

Quentin Bolliet, Aymeric Floyrac**, Sophie Maillard*** et Agathe Rosenzweig*****

Le décideur montre un intérêt croissant pour la prise de décision basée sur les données. L'administration, de son côté, dispose d'une grande quantité de données, d'infrastructures et de ressources humaines capables de les exploiter. Ces développements récents permettent aux inspections, comme l'inspection générale des finances (IGF), de renforcer leurs missions d'évaluation et de conseil par des études spécifiques mobilisant des données déjà disponibles. L'IGF utilise celles-ci pour quantifier les constats et consolider les analyses. À la différence du système statistique public, cependant, elle ne produit pas de données et n'en assure pas le suivi dans le temps.

À titre d'exemple, une mission réalisée par le pôle science des données de l'IGF sur l'assurabilité des biens des collectivités territoriales est présentée. Elle illustre comment une utilisation large des données de l'Institut national de l'information géographique et forestière (IGN) et de la Direction générale des Finances publiques (DGFiP) permet, grâce aux techniques de *data science*, d'éclairer de manière très fine un questionnement adressé à l'IGF.

Si la demande pour des études sur mesure augmente et que le modèle d'un pôle dédié à la science des données essaime dans les inspections, il comporte des conditions de réussite, tant sur la culture de l'analyse quantitative que sur la qualité des données.

Decision-makers are showing a growing interest for data-driven decision making. At the same time, the administration has access to vast amounts of data, infrastructures and skilled human resources to leverage them. These recent developments allow inspectorates, such as the General Inspectorate of Finances (IGF), to strengthen their evaluation and advisory missions through specific studies using already available data. The IGF uses this data to quantify its observations and strengthen its analyses. Unlike the official statistics system, however, it neither produces data nor monitors it over time.

One example is a study conducted by the IGF's data science team about the insurability of municipalities' real estate assets. It demonstrates how combining data from the National Institute of Geographic and Forestry Information (IGN) and the Directorate General of Public Finances (DGFiP) with data science techniques can provide highly detailed insights to address policy questions submitted to the IGF .

Although demand for tailored studies is increasing and other inspectorates have been replicating the model of a data science dedicated hub, this model needs to fulfil conditions relating both to the culture of quantitative analysis and the quality of the data.

* *Data scientist*, inspection générale des finances.
quentin.bolliet@igf.finances.gouv.fr

** Responsable adjoint du pôle science des données, inspection générale des finances.
aymeric.floyrac@igf.finances.gouv.fr

*** Responsable du pôle science des données, inspection générale des finances.
sophie.maillard@igf.finances.gouv.fr

**** *Data scientist*, inspection générale des finances.
agathe.rosenzweig@igf.finances.gouv.fr

> ***L'inspection générale des finances (IGF), forte de sa tradition d'objectivation et d'évaluation des politiques publiques, s'est dotée en 2019 d'un pôle science des données, qui compte aujourd'hui une dizaine d'agents.***

La prise de décision, qu'elle soit le fait des administrations, des entreprises ou des individus, repose de plus en plus sur l'exploitation des données. En parallèle, la quantité de données disponibles a explosé au cours des deux dernières décennies, tandis que les techniques de traitement, tant statistiques qu'informatiques, se sont démocratisées et sont aujourd'hui bien plus accessibles qu'auparavant. Cette évolution a été soutenue par le développement d'infrastructures robustes, capables de gérer des traitements parfois massifs : les plateformes de *data science*.

Sous réserve d'investir dans les ressources matérielles et surtout humaines nécessaires, ces transformations ouvrent la possibilité aux acteurs publics et privés de réaliser des analyses quantitatives mieux ciblées sur leurs sujets d'intérêt et questionnements spécifiques. En particulier, elles offrent aux acteurs publics une opportunité unique d'élaborer des politiques mieux informées, fondées sur des preuves tangibles, afin de répondre aux besoins des citoyens de manière plus efficace et équitable.

L'inspection générale des finances (IGF), forte de sa tradition d'objectivation et d'évaluation des politiques publiques (Prada, 2012), s'est ainsi dotée en 2019 d'un pôle science des données, qui compte aujourd'hui une dizaine d'agents, capable de mener à bien des travaux à forte composante quantitative au service de ses missions d'évaluation et de conseil.

La science des données à l'IGF : une réponse moderne à un besoin ancien

Les institutions publiques s'appuient depuis longtemps sur le système statistique public (SSP), qui offre des productions générales de haute qualité pour éclairer les politiques publiques. Cependant, ces productions ne permettent pas de répondre à toutes les questions que se pose le décideur : les données existantes, souvent conçues pour des objectifs larges, peuvent ne pas répondre précisément aux besoins spécifiques et parfois étroits d'une mission d'évaluation ou de conseil. Par ailleurs, les délais imposés à la décision publique – et par conséquent à certaines missions, souvent de quelques mois – rendent impossible la création de nouvelles données adaptées.

Depuis la loi pour une République numérique de 2016[1] et le rapport Bothorel (Bothorel, 2020), l'augmentation du volume de données administratives disponibles a ouvert de nouvelles perspectives pour une exploitation ad hoc. Néanmoins, pour en tirer pleinement parti, les institutions doivent disposer d'équipes dédiées et compétentes en analyse de données. C'est la raison d'être du pôle science des données de l'IGF, comme des pôles équivalents créés dans d'autres institutions, notamment à la Cour des comptes et dans

1 Voir les références juridiques en fin d'article.

d'autres inspections générales (voir dernière partie). Ces pôles travaillent en interaction permanente avec des profils généralistes (que sont par exemple les inspecteurs des finances), pour répondre à des questions variées : objectiver une situation à partir de données existantes, évaluer les effets de politiques publiques, ou simuler les impacts potentiels d'une réforme. On peut par exemple citer la mission sur les ressources humaines de l'État dans le numérique, le comité d'évaluation du plan France Relance ou encore la mission sur le coût du travail en France, portée par Antoine Bozio et Étienne Wasmer.

Mobiliser les données pour répondre à des questions de politique publique

Les pôles science des données de l'IGF et des autres instances d'évaluation sont composés de statisticiens et de *data scientists* (Comte et al., 2022). Ils se distinguent de par leur rôle des structures du système statistique public. Contrairement à l'Insee et aux services statistiques ministériels, ils n'ont pas vocation à produire des statistiques générales, périodiques et comparables. Ainsi, ils ne participent pas à « la production statistique d'informations économiques ou sociales », domaine qui délimite selon Michel Volle la « statistique » telle qu'entendue dans son acception courante (Volle, 1984). Ils se concentrent sur des analyses ponctuelles et ciblées, adaptées aux besoins à court terme des décideurs. À ce titre, leur hiérarchie des valeurs diffère de celle du système statistique public :

1. **Réactivité** : contraints par les délais des missions, les pôles doivent souvent adapter au temps qui leur est imparti le degré de finesse de leurs analyses : un résultat produit hors délai ne sera jamais communiqué au commanditaire de la mission.

2. **Adaptabilité** : ces pôles travaillent avec des données issues de multiples sources administratives, souvent hétérogènes. Cela nécessite des infrastructures partagées et sécurisées pour simplifier l'accès aux données et garantir leur confidentialité, tout en réduisant le temps consacré à leur préparation.

3. **Transparence et comparabilité** : ces pôles n'ont pas pour objectif de produire des données publiques ou comparables entre pays, contrairement à la statistique publique (de Peretti et Touchelay, 2024). Leur priorité est l'aide à la décision, souvent dans un cadre confidentiel. Ainsi, dans le cas de l'IGF, il appartient au commanditaire de la mission de décider du caractère public ou non du rapport[2].

> ***Les analyses produites par les pôles de science des données apportent une réponse moderne à un besoin ancien : produire des analyses sur mesure, rapidement mobilisables, pour répondre aux défis immédiats des politiques publiques.***

Ces nouvelles pratiques rappellent les origines mêmes de la statistique, pensée initialement comme un outil au service de l'État pour éclairer et guider l'action publique. Elles répondent au besoin très ancien des « pratiques de mise en chiffre », dont les objectifs sont « de coordonner les activités humaines, de trouver des accords entre parties prenantes, de prendre et de justifier des décisions fiscales, militaires ou sociales » (Martin, 2023). Ainsi, à l'image des enquêtes

2 Le commanditaire de la mission est généralement un membre du gouvernement, le plus souvent le Premier ministre ou le ministre de l'Économie et des Finances.

ponctuelles de Colbert ou des dénombrements de Vauban, les analyses produites par les pôles de science des données s'inscrivent dans une tradition d'investigation spécifique et orientée.

Cependant, ces nouvelles pratiques s'écartent des standards actuels de la statistique publique, qui met particulièrement en avant les valeurs de transparence et de comparabilité. En contrepartie, elles apportent une réponse moderne à un besoin ancien : produire des analyses sur mesure, rapidement mobilisables, pour répondre aux défis immédiats des politiques publiques (***figure 1***).

Exploiter une variété de sources issues, entre autres, de la statistique publique

Le pôle science des données de l'IGF bénéficie pour la conduite de ses travaux du mouvement général de décloisonnement des données publiques et de la diffusion des outils permettant de les exploiter. Les questions spécifiques traitées par les missions le conduisent à manipuler une grande diversité de données :

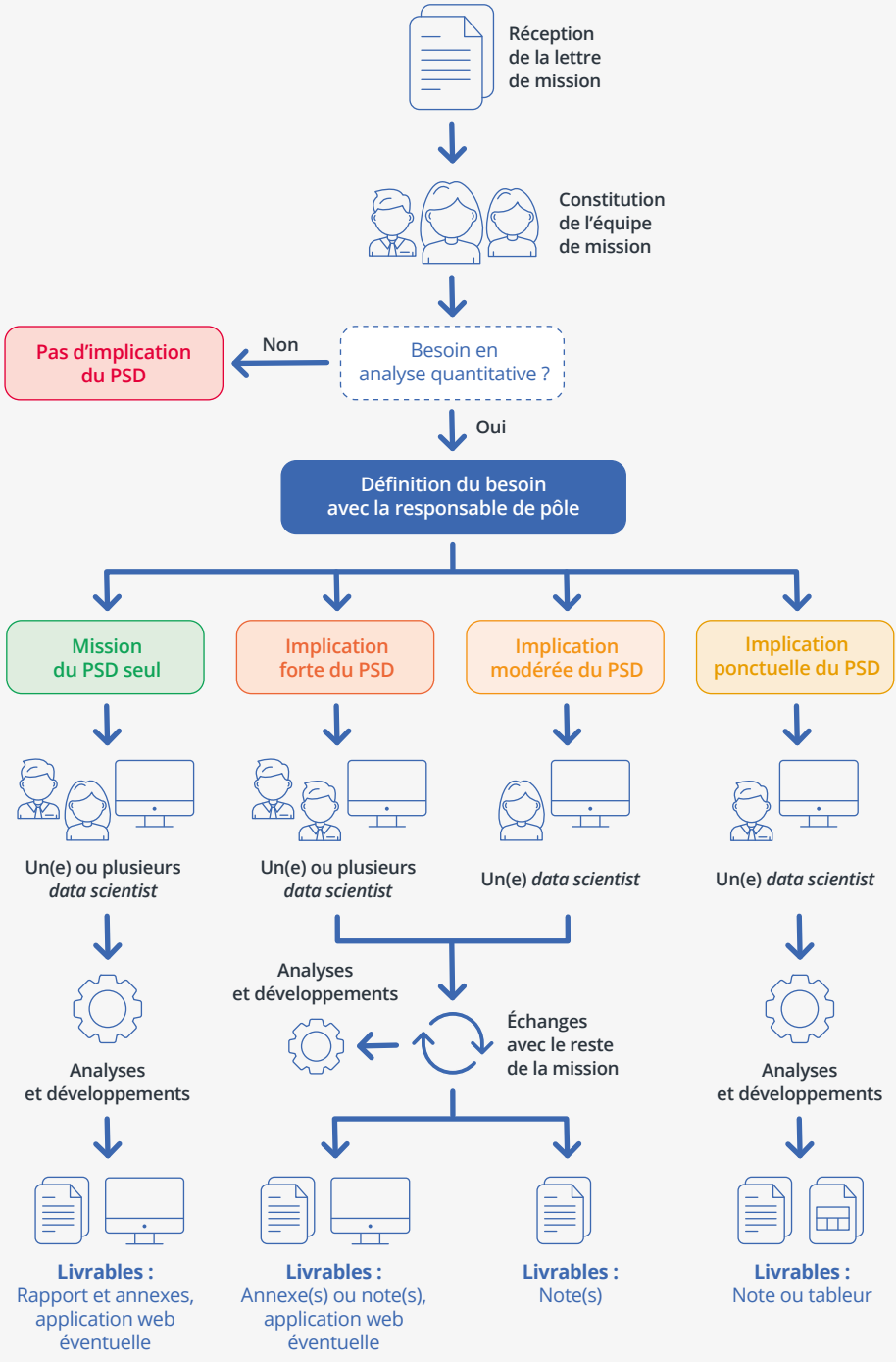
- **des données agrégées produites par la statistique publique** : comme tout citoyen, les membres du pôle ont accès à ces données dont la qualité est garantie ;

- **des données individuelles produites par la statistique publique** : les membres du pôle sont habilités par le comité du secret statistique à accéder à un large catalogue de ces sources via le centre d'accès sécurisé aux données (CASD) [Gadouche, 2019], dont notamment les données d'imposition des individus (fichier POTE[3]) et des entreprises (fichiers FARE[4] et BIC-IS[5]), ainsi que celles issues des déclarations sociales nominatives (DSN) des employeurs. Ces accès s'accompagnent d'obligations strictes en matière de respect du secret statistique, avec une distinction claire entre les prérogatives des statisticiens du pôle, habilités à utiliser ces données, et celles des inspecteurs des finances, qui ne peuvent accéder qu'aux résultats agrégés des analyses ;

- **des données d'autres sources que celles de la statistique publique** : au gré des thématiques abordées par les missions, les membres du pôle peuvent aussi mobiliser, au-delà de cet ensemble déjà riche, d'autres données souvent moins structurées, ou se prêtant moins naturellement à des traitements statistiques :
  - des données textuelles, par exemple issues des bulletins officiels ;
  - des données géographiques, notamment produites par l'Institut national de l'information géographique et forestière (IGN), comme la BD Forêt ou la BD TOPO, mais aussi les données cadastrales ;
  - et, naturellement, des données de gestion des administrations ou autres organismes rencontrés par les inspecteurs. La manipulation de celles-ci demande une certaine prudence ; il est en effet nécessaire de s'assurer de leur qualité et de leur périmètre, mais aussi de la cohérence des concepts qui les sous-tendent avec ceux nécessaires à la mission d'inspection. Il peut s'agir par exemple de données issues de logiciels internes, qui seraient remplies à la main (avec le risque d'erreur et d'inconsistance que

3 Le fichier POTE rassemble les informations recueillies dans le cadre des déclarations des foyers fiscaux à l'impôt sur le revenu des personnes physiques (IRPP).
4 Le fichier FARE rassemble les données individuelles comptables des entreprises.
5 Le fichier BIC-IS rassemble les informations recueillies dans le cadre des déclarations de bénéfices industriels et commerciaux (BIC), pour différents régimes d'imposition, et des déclarations à l'impôt sur les sociétés (IS).

Figure 1 - Fonctionnement du pôle science des données de l'inspection générale des finances



PSD : Pôle science des données.

cela entraîne), sous un format potentiellement propriétaire, sans identifiant commun avec d'autres bases accessibles par ailleurs, et donc sans possibilité d'appariement (Koumarianos et al., 2024).

Parfois, il arrive que le premier niveau de données soit suffisant pour la mission, auquel cas le rôle du pôle se borne à accompagner les inspecteurs dans la compréhension des concepts statistiques sous-jacents aux données en question, en expliquant par exemple le fonctionnement de la nomenclature d'activité française (NAF)[6], ce qu'est le numéro Siren[7], etc.

Souvent, les données existantes ne permettent de répondre aux questions que se posent les missions que de façon très imparfaite : elles peuvent être trop anciennes, ne couvrir que partiellement le champ d'intérêt ou, pour certaines données de gestion, souffrir de défauts dans leur conception tels que les résultats obtenus doivent être considérés avec grande précaution. Ce dernier type de données « externes non maîtrisées » met aussi le statisticien-*data scientist* au défi d'entrer dans la logique de constitution des fichiers afin de « reconstituer a posteriori un pseudo-appareil d'observation, à partir de données qui n'ont pas été conçues à cette fin » (Rivière, 2020). Cela passe par des échanges étroits avec les producteurs, gestionnaires ou utilisateurs de ces fichiers tout au long des travaux.

> ***La démarche initiale du* data scientist *est systématiquement la même : comprendre le sens des variables, les clés entre les fichiers, puis essayer de reproduire, quand ils existent, des résultats agrégés produits par une autre administration.***

Quelle que soit la catégorie des données mobilisées, la démarche initiale du *data scientist* est systématiquement la même : comprendre le sens des variables, les clés entre les fichiers, puis essayer de reproduire, quand ils existent, des résultats agrégés produits par une autre administration. Il s'agira par exemple de retrouver le montant total d'exonération de cotisations sociales, ou bien le nombre total de patients en affection de longue durée en France hexagonale. Cette étape est cruciale et permet de s'assurer que le *data scientist* a compris la signification des variables et qu'il applique les bons filtres et transformations à sa base de données. Une fois les faits stylisés établis, le *data scientist* peut sereinement mener son analyse pour répondre aux questions réelles de la mission. Bien sûr, dans le cas de données issues du CASD, le pôle s'enrichit de ses bonnes pratiques acquises au fur et à mesure des missions, si bien qu'il a développé des traitements de données standardisés (ou *pipelines*) pour certaines bases fréquemment utilisées, ce qui permet de limiter le risque d'erreur.

Utiliser des infrastructures de calcul développées par l'écosystème

Le pôle est un utilisateur enthousiaste du CASD qui, en sus d'héberger les données sensibles, offre également des serveurs de calcul permettant de réaliser des analyses dans une bulle sécurisée. Pour traiter les données hors CASD, en revanche, il est nécessaire

6 *https://www.insee.fr/fr/information/2406147.*
7 *https://www.insee.fr/fr/metadonnees/definition/c2047.*

de disposer d'un environnement adapté : les ordinateurs du ministère n'ont encore pas la puissance de calcul nécessaire pour traiter des données volumineuses, par exemple des données géographiques fines, et les politiques de sécurité informatique brident largement l'installation de logiciels nécessaires à ces traitements.

Pour contourner ces difficultés, le pôle a recours à deux plateformes de *data science* offrant un vaste catalogue de services : le SSPCloud et Nubonyxia. Ces deux plateformes sont des instances d'Onyxia (Comte et al., 2022), une solution de *data science* intégralement développée par l'Insee et accessible en *open source*, qui présente une souplesse, une puissance et une ergonomie parfaitement adaptées aux besoins du pôle. Le SSPCloud, développé et hébergé par l'Insee, est la plateforme la plus mature, mais sa principale limite est de ne pas permettre d'héberger des données confidentielles. Nubonyxia, de son côté, est un déploiement d'Onyxia sur les serveurs Nubo de la Direction générale des Finances publiques (DGFiP) réalisé par le Bercy Hub. Ce projet est à un stade de développement plus précoce, mais il offre des garanties de sécurité plus importantes et est connecté au réseau interministériel de l'État. Ces deux infrastructures disposent par ailleurs de tous les services nécessaires à la plupart des projets de *data science* : environnement de développement intégré (VSCode, Rstudio, JupyterLab), système de gestion de base de données (PostgreSQL), suivi d'entraînement de modèle de *machine learning* (MLFlow). Elles permettent également de déployer des applications web développées en Python ou en R, et de les mettre à disposition des autres membres des missions ou des commanditaires.

Malgré les qualités de ces plateformes, il peut cependant arriver que, dans un arbitrage entre la qualité du rendu et les délais que le pôle doit tenir, les exploitations de données n'aillent pas jusqu'à leur terme théorique. Ainsi, dans le cadre d'une mission sur la gestion de la forêt privée, le *data scientist* a dû se limiter à travailler sur un échantillon de dix départements forestiers pour produire ses résultats, les volumes de données à traiter étant trop importants (de l'ordre d'une dizaine de gigaoctets par département).

Implémenter des modèles de *data science* pour objectiver les faits

Les missions de l'IGF auxquelles le pôle est associé sont essentiellement des missions d'évaluation des politiques publiques. Dans ce contexte, la plus-value du pôle science des données est sa polyvalence, c'est-à-dire sa capacité à mobiliser des techniques quantitatives variées, et notamment aussi bien des méthodes économétriques que des méthodes de *machine learning*.

> ***La plus-value du pôle science des données est sa polyvalence, c'est-à-dire sa capacité à mobiliser des techniques quantitatives variées, et notamment aussi bien des méthodes économétriques que des méthodes de* machine learning.**

Ces missions comprennent généralement deux phases distinctes : une phase de diagnostic et une phase de simulation de réforme. Pour chacune d'elles, il est crucial de procéder avec des méthodes scientifiques et notamment de **s'assurer de raisonner « toutes choses égales par ailleurs »**. Les délais courts (2 mois en général) dans lesquels les conclusions des travaux du pôle sont attendues exigent de procéder avec beaucoup d'efficacité, d'abord

pour comprendre et pré-traiter les données (nettoyage, changements de nomenclatures, traitement des valeurs manquantes), puis pour construire un contrefactuel convaincant, écarter les variables confondantes[8], estimer le modèle lui-même et, enfin, l'interpréter (Givord, 2014). Quand la tâche impose de construire un modèle de microsimulation (Blanchet, 2020), la même efficacité est requise. La capitalisation sur les codes produits d'une mission à l'autre se révèle ainsi essentielle pour livrer les productions à temps et harmoniser les réalisations du pôle.

Les méthodes de *machine learning* ont des usages plus variés : prédire des valorisations, dégager des regroupements (ou *clusters*) d'individus, faire du codage automatique, etc. Au-delà des distinctions méthodologiques entre les tâches confiées au pôle, un point commun à celles-ci est que les *data scientists* ont vocation à produire une méthodologie ou une application qui reste à l'état de « preuve de concept ». En effet, contrairement au système statistique public, l'IGF n'a pas vocation à réaliser des productions périodiques : les modélisations, évaluations et chiffrages sont faits dans le cadre d'une mission délimitée dans le temps et ne sont pas destinés à être répliqués d'une année à l'autre. Si toutefois cette réplication s'avère utile pour le décideur, il appartient aux administrations compétentes de reprendre à leur compte les méthodologies développées par le pôle, de les améliorer et de les intégrer à leur production statistique. En particulier, selon les cas et le statut de confidentialité de la mission, les codes produits peuvent être directement transmis aux services concernés, ou mis en ligne pour contribuer à des projets en code source ouvert.

▶ Un exemple de mission : l'assurabilité des collectivités territoriales

Afin d'illustrer de façon plus concrète le type de travaux conduits, nous présentons le soutien apporté par l'équipe du pôle science des données à une mission portant sur l'assurabilité des biens des collectivités locales, mission à laquelle l'IGF et l'inspection générale de l'administration ont apporté une assistance technique[9].

Attendu : objectiver les difficultés des collectivités locales à assurer leurs biens

La mission a été lancée à la suite de difficultés financières rencontrées par certains acteurs de l'assurance des biens des collectivités territoriales et d'obstacles, signalés par certaines communes, au renouvellement de leur contrat d'assurance. Ces difficultés et obstacles s'inscrivent dans un contexte d'accroissement des risques climatiques et après les épisodes de violences urbaines à l'été 2023, ayant engendré plus d'un milliard d'euros de dommages. Dans le cadre de cette mission, les membres du pôle science des données ont contribué à quantifier l'évolution des dépenses d'assurance des collectivités territoriales, à mettre en évidence leurs déterminants et à estimer la valeur du patrimoine exposé à un ou plusieurs risques naturels.

8 Variables qui influencent à la fois une variable dépendante et une variable indépendante.
9 Mission d'assistance auprès des personnalités qualifiées Alain Chrétien, maire de Vesoul, et Jean-Yves Dagès, ancien président de Groupama.

> ***Les membres du pôle science des données ont contribué à quantifier l'évolution des dépenses d'assurance des collectivités territoriales, à mettre en évidence leurs déterminants et à estimer la valeur du patrimoine exposé à un ou plusieurs risques naturels.***

Ces estimations relatives à l'exposition du patrimoine des collectivités territoriales font écho à des travaux menés par des acteurs de la statistique publique sur le patrimoine immobilier des ménages à partir du répertoire Fidéli[10] (André et Meslin, 2022 et 2025). Cependant, ces travaux diffèrent par le champ sur lequel ils portent : les logements possédés par les ménages dans le cas des travaux conduits par l'Insee, les biens des communes et communautés de communes pour les travaux de la mission. Ces derniers ont aussi été menés dans un délai serré de quelques mois, ce qui induit des différences importantes dans le retraitement des données et les méthodes mises en œuvre.

Données : inventorier le patrimoine des communes à l'aide du cadastre

Le premier chantier entrepris afin de caractériser le patrimoine immobilier et son exposition aux risques naturels a consisté en l'inventaire des biens immobiliers possédés par les communes. En effet, si un tel inventaire existe pour les ménages (Fidélimmo[11] côté Insee ou Majic[12] côté DGFiP), ce n'est pas le cas pour les collectivités territoriales[13]. Or, il était indispensable de pouvoir localiser finement tous les bâtiments possédés par les communes (dont les écoles, les églises, etc.), afin de mesurer leur exposition aux risques naturels. Il fallait aussi pouvoir valoriser chaque bâtiment et donc recueillir des caractéristiques sur lesquelles s'appuyer pour estimer une valeur vénale. En lien avec la sous-direction de la gestion comptable et financière des collectivités territoriales et le bureau du cadastre de la DGFiP, une base de données inventoriant et valorisant le patrimoine bâti des collectivités territoriales a ainsi été construite.

Les parcelles appartenant à des communes ou des établissements publics de coopération intercommunale (EPCI) sont recensées à partir des fichiers des locaux et des parcelles des personnes morales, dérivés des données cadastrales et publiés par la DGFiP. Ainsi, 356 000 parcelles bâties et 5,4 millions de parcelles non bâties appartenant à une collectivité du bloc communal[14] en 2023 ont pu être approchées à partir de ces données. Les infrastructures telles que les ouvrages d'art et de voirie n'y sont cependant pas référencées.

10 *https://www.insee.fr/fr/metadonnees/source/serie/s1019.*
11 Fidélimmo est une extension du répertoire Fidéli qui vise à mieux analyser le patrimoine immobilier des ménages.
12 Le fichier Majic rassemble les informations cadastrales sur les parcelles, bâties ou non bâties, et les locaux (propriétés bâties).
13 Les données comptables renseignent sur la valorisation du patrimoine détenu par les collectivités, mais ne sont pas détaillées bien par bien.
14 On désigne par là une commune, une communauté de communes, une communauté de villes, une communauté d'agglomérations, une communauté urbaine, un syndicat intercommunal à vocation multiple (SIVOM) ou un syndicat intercommunal à vocation unique (SIVU).

Cet inventaire a ensuite été enrichi à partir d'autres sources de données géocodées :

- La table « Bâti – Parcelle » de la BAN PLUS, produite par l'IGN, associe à chaque parcelle référencée les bâtiments qu'elle abrite : 535 000 bâtiments ont ainsi pu être identifiés sur 224 000 des 356 000 parcelles bâties (en ne retenant que la parcelle principale de chaque bâtiment) ;

- La table « Bâtiment » de la BD TOPO, produite par l'IGN, fournit les caractéristiques physiques des bâtiments (surface au sol, hauteur, état) ;

- La table « ERP » de la BD TOPO renseigne sur la nature et la capacité d'accueil des établissements recevant du public (ERP) comme les mairies, églises, écoles, etc. ;

- La base permanente des équipements[15] (BPE), produite par l'Insee et géocodée, répertorie en outre un large éventail d'équipements et de services, marchands ou non, accessibles au public au 1er janvier de chaque année. Elle renseigne ainsi sur la fonction des bâtiments publics dont les collectivités sont propriétaires. En 2021, elle porte sur 188 types de services et équipements différents, répartis en sept grands domaines : services aux particuliers, commerces, enseignement, santé-social, transports-déplacements, sports-loisirs-culture et tourisme (Helfenstein, 2022).

L'appariement entre les données cadastrales et la table « Bâtiment » est réalisé en utilisant l'identifiant cadastral de la parcelle. Les informations contenues dans la table « ERP » et dans la BPE sont ajoutées à la base de données en superposant les informations géocodées dans ces deux sources avec la géométrie des bâtiments et des parcelles (***figure 2***).

Modélisation : estimer la valeur du patrimoine par apprentissage

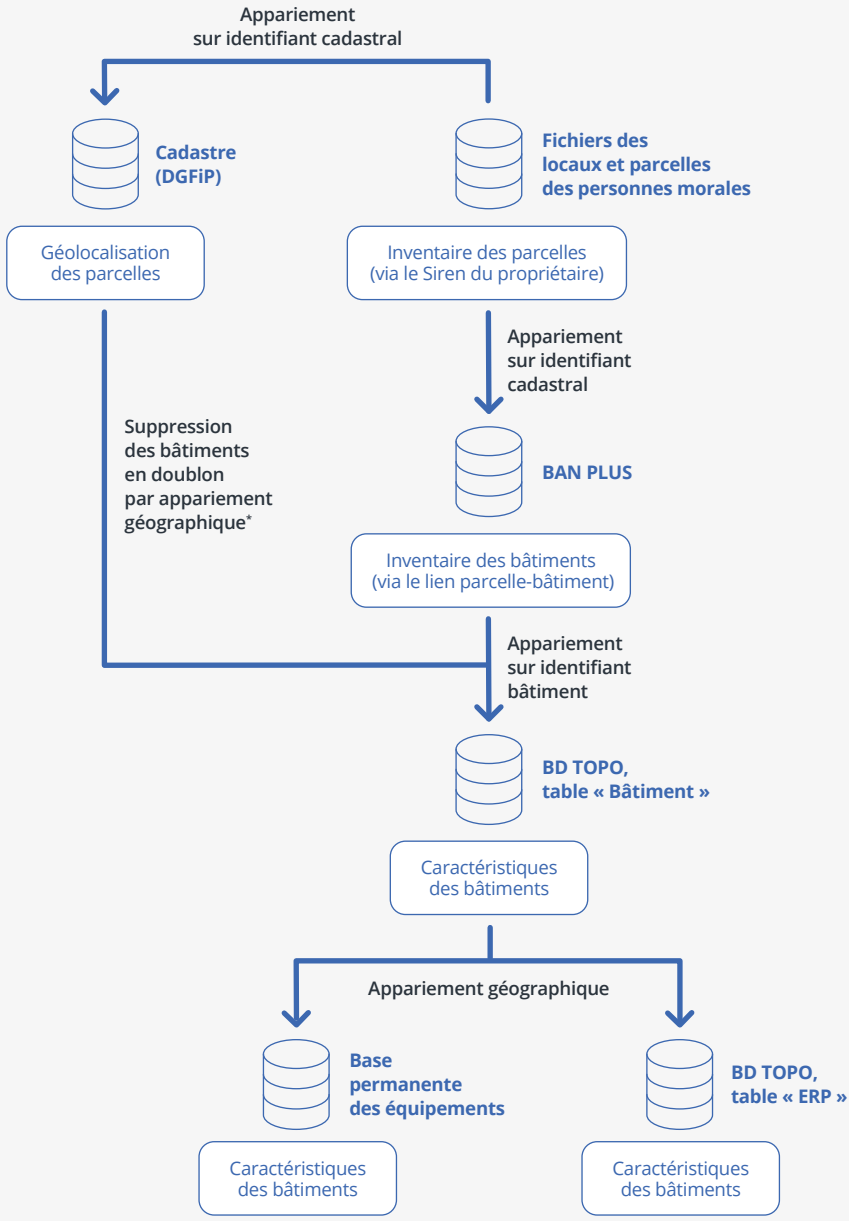
Pour mesurer l'exposition du patrimoine des collectivités aux risques, il est nécessaire de quantifier ce patrimoine, en l'occurrence via une valeur vénale. Cette valeur n'est naturellement pas présente dans les données citées ci-dessus. Il faut donc trouver une façon d'en construire une estimation, typiquement à l'aide d'un modèle de *machine learning* qui apprendrait la valeur en fonction d'un certain nombre de caractéristiques. Cela suppose d'avoir une base de référence qui présente à la fois les caractéristiques en question et une valeur vénale, c'est-à-dire un prix de vente.

L'inférence de la valeur vénale des biens des collectivités a donc été réalisée à partir du fichier des demandes de valeurs foncières (DVF) publié par la DGFiP. Ce dernier recense l'ensemble des ventes de biens fonciers réalisées entre 2018 et 2023 en France, sauf à Mayotte et en Alsace-Moselle, soit 7,7 millions de transactions sur cinq ans. La base contient en particulier les prix de mutation enregistrés par les notaires. Toutefois, il y a peu de ventes de biens publics dans ces données. L'estimation de la valeur des biens des collectivités a donc été conduite en deux étapes : une première qui permet d'obtenir le prix du bien s'il était initialement détenu par un acteur privé et une deuxième qui adapte ce prix au fait que le bien serait hypothétiquement vendu par un acteur public :

- Un premier modèle XGBoost (***encadré 1***) est entraîné sur les 3,1 millions de transactions de biens n'appartenant pas à des collectivités en 2023 pour lesquelles les prix et les

15 *https://www.insee.fr/fr/metadonnees/source/serie/s1161.*

Figure 2 - Appariement de diverses sources de données permettant d'inventorier et caractériser le patrimoine des collectivités territoriales



*Certains bâtiments sont à cheval sur plusieurs parcelles et apparaissent donc plusieurs fois dans la base, liés à plusieurs parcelles. Un bâtiment dont 75 % de la géométrie se trouve dans la parcelle A et 25 % dans la parcelle B sera rattaché à la parcelle A.
BAN PLUS : Base Adresse Nationale PLUS.
BD TOPO : Base de données topographiques.
ERP : Établissements recevant du public.
DGFiP : Direction générale des Finances publiques.

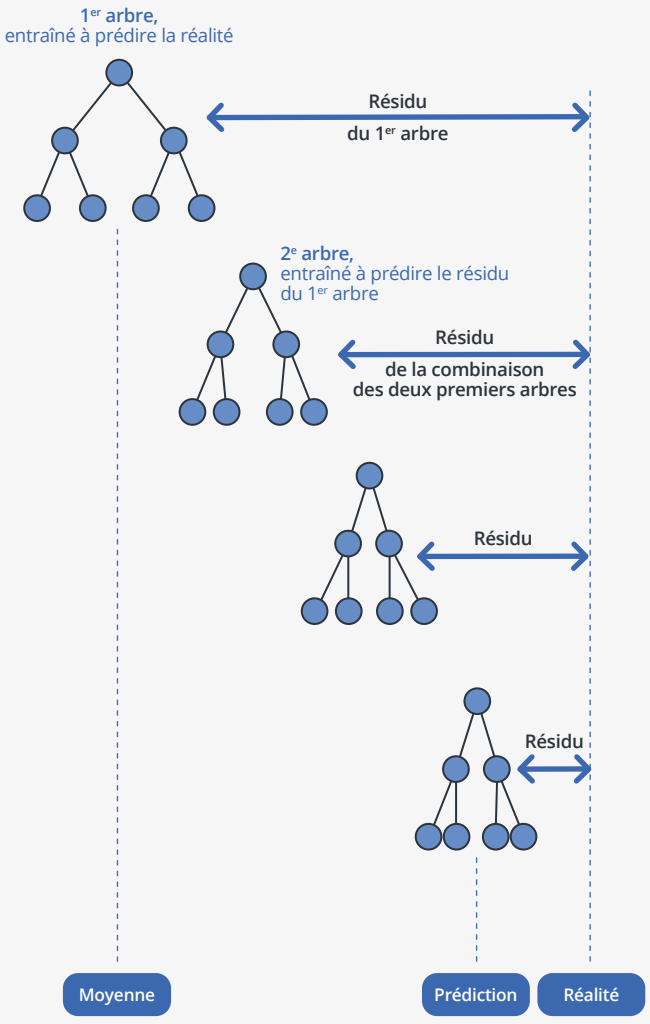
▶ Encadré 1. Fonctionnement de l'algorithme XGBoost

L'algorithme XGBoost (*Extreme Gradient Boosting Tree*) est un modèle d'apprentissage basé sur des arbres de régression ou de classification (Chen et Guestrin, 2016). La statistique publique l'utilise dans certains travaux (André et Meslin, 2025).

L'algorithme d'entraînement combine les prédictions de nombreux modèles simples pour créer un modèle plus puissant. Pour réaliser cette combinaison, il emploie des méthodes dites de *boosting* (Freund et Schapire, 1996), c'est-à-dire que le modèle apprend en construisant itérativement des arbres de décision, chaque arbre étant entraîné à prédire les erreurs de prédiction des arbres précédents combinés (***figure encadré***).

Ainsi, le premier arbre est entraîné à prédire la réalité. Puis, le deuxième arbre est entraîné à prédire les résidus du premier arbre. Les prédictions du deuxième arbre sont alors multipliées par un facteur η inférieur à 1 et ajoutées aux prédictions du premier arbre. Un troisième arbre est alors entraîné à prédire les résidus de la combinaison des deux premiers arbres, écart entre la réalité et la combinaison des prédictions, etc.

L'algorithme XGBoost est optimisé, notamment pour le calcul parallèle, afin de pouvoir être entraîné rapidement, et régularisé pour empêcher que le modèle devienne trop complexe, diminuant ainsi les risques de surapprentissage.



Le modèle final est issu de la combinaison des arbres successifs.

caractéristiques des bâtiments sont correctement renseignés. Le modèle s'appuie sur les caractéristiques physiques du bâtiment (nature, surface, hauteur, nombre d'étages, superficie de la parcelle) et sur des informations afférentes à son environnement (coordonnées, revenu médian de l'Iris[16], caractéristiques de la commune) pour estimer un prix de vente.

- Un second modèle XGBoost est ensuite entraîné sur les 43 000 ventes de parcelles appartenant en 2023 aux communes et aux EPCI à partir des mêmes caractéristiques que celles utilisées en première approche, enrichies des prédictions du premier modèle. Cette deuxième étape permet de raffiner le modèle sur des données plus spécifiques à notre cas d'usage, tout en exploitant les résultats du premier modèle.

Suivant cette approche, le patrimoine immobilier des communes et EPCI s'élèverait à 428 milliards d'euros (Md€), soit une valeur de même ordre mais supérieure aux estimations tirées des données comptables (350 Md€) ou de la comptabilité nationale (275 Md€). Ces trois décomptes souffrent de limites qui leur sont propres. Pour celui mis en œuvre ici, la difficulté est d'extrapoler la valeur des biens vendus, qui peuvent être assez spécifiques, aux biens possédés par les collectivités n'ayant souvent pas vocation à être cédés (écoles, églises, piscines, etc.). Si les valeurs de patrimoine agrégées diffèrent, la répartition géographique du patrimoine est cependant assez cohérente dans notre approche et dans les données comptables.

Dernière étape : croiser l'estimation du patrimoine et les risques naturels

Une fois les bâtiments inventoriés, caractérisés et associés à une valeur vénale approchée, l'objectif est d'identifier les différents risques naturels pesant sur chacun d'eux (***encadré 2***). Le choix des bases de données permettant de qualifier le niveau des risques encourus a été effectué en suivant les recommandations de différents producteurs de données rencontrés par la mission, notamment le Service des données et études statistiques (SDES)[17], le Cerema[18] et la Direction générale de la prévision des risques (DGPR). Les cartes de risques au niveau infracommunal ont été retenues lorsque les données existaient, ce qui est le cas pour le retrait-gonflement des sols argileux, les inondations, via l'enveloppe approchée du risque d'inondation potentiel (EAIP), et les feux de forêt. Dans le cas contraire, des données mises en ligne par l'Observatoire national sur les effets du réchauffement climatique (ONERC) et produites par le SDES ont été utilisées pour mesurer les risques à l'échelle communale (mouvements de terrain, séismes).

Cette approche présente cependant des limites inhérentes aux données disponibles pour documenter l'exposition aux différents risques du territoire français. Ainsi, les mesures peuvent se révéler imprécises. En outre, ces données ne permettent pas de caractériser la sinistralité induite par la matérialisation d'un risque donné.

L'exposition du patrimoine des communes est finalement qualifiée par la superposition de l'inventaire du patrimoine total et des cartes des risques naturels ainsi construites (***figure 3***). 240 000 bâtiments des communes sont situés dans des zones exposées à un

16 *https://www.insee.fr/fr/metadonnees/definition/c1523*.
17 Service statistique ministériel de l'énergie, du logement, du transport et de l'environnement.
18 Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement.

▶ Encadré 2. Les risques naturels retenus dans l'analyse

Certains risques naturels ont été écartés de l'analyse : le recul du trait de côte, qui est jugé trop déterministe pour être pris en compte dans un mécanisme assurantiel, et le risque de grêle, dont l'emplacement géographique est trop difficile à prédire. Les risques pris en compte sont les suivants :

• **Le risque d'inondation** est mesuré par l'enveloppe approchée du risque d'inondation potentiel (EAIP). Définie à l'échelle infracommunale, l'EAIP permet de tenir compte de deux aléas hydrologiques distincts : les inondations par submersion marine et les débordements de cours d'eau ou de crues. Elle renseigne sur le risque potentiel, sans détailler l'intensité du risque afférent aux différentes zones concernées. Son approche est maximaliste au regard de l'étendue des zones de risques.

• **Le risque de retrait-gonflement des sols argileux** (RGA) est documenté par une cartographie départementale réalisée entre 1997 et 2010 par le Bureau de recherches géologiques et minières (BRGM). Ce risque, lié aux conditions météorologiques, n'est pas dangereux pour l'humain, mais est potentiellement très dommageable pour le bâti.

• **Les risques de mouvement de terrain et de séisme** sont cartographiés à la maille communale à partir de données mises en ligne par l'Observatoire national sur les effets du réchauffement climatique (ONERC). Ils incluent notamment les glissements de terrain, les coulées boueuses de diverses intensités, les affaissements et effondrements dus à des cavités souterraines.

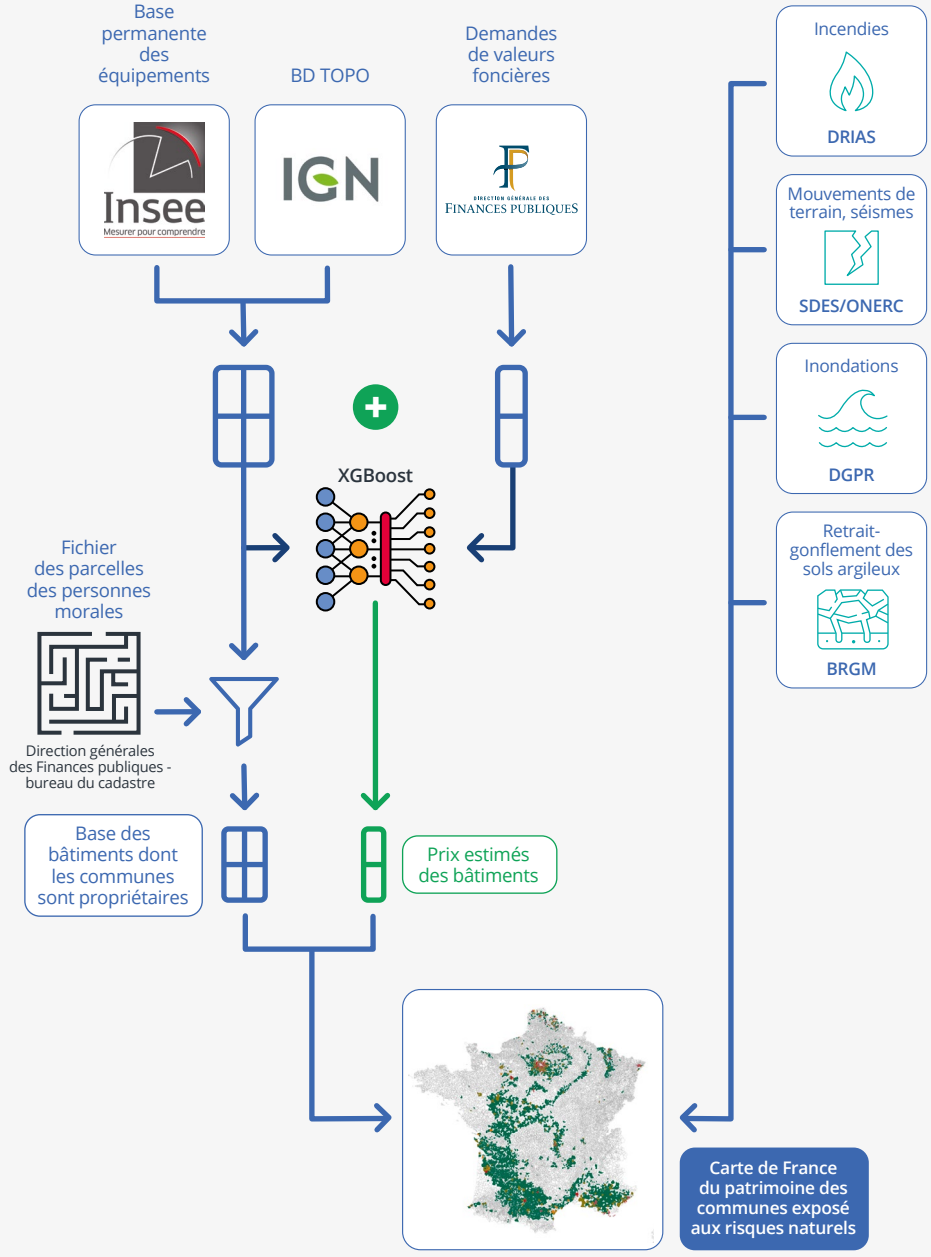
• **Le risque de feux de forêt** est quantifié par l'indice Feu Météo (IFM), qui rend compte de la propension à l'éclosion et à la propagation des feux. Cet indice peut être calculé à des horizons temporels éloignés, en s'appuyant sur les scénarios climatiques du deuxième Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC2). Le scénario retenu est le plus proche de la trajectoire de référence pour l'adaptation au changement climatique (TRACC), soit le scénario RCP 4.5, à partir duquel est calculé un indice Feu météo journalier. Le nombre de jours tels que l'indice Feu Météo est supérieur au seuil de 40 est retenu comme indicateur de risque de feux de forêt à l'horizon 2100.

risque sévère et géolocalisé précisément, soit 49 % des bâtiments des communes et 51 % de leur patrimoine bâti estimé par les données de transactions immobilières. Le facteur de risque sévère précisément appréhendé par les données et qui touche le plus de biens correspond aux inondations provoquées par les débordements de cours d'eau (31 % des bâtiments et 40 % du patrimoine bâti). La part des bâtiments des communes exposée aux risques est proche de celle des personnes morales, pour lesquelles l'analyse a été répliquée à titre de comparaison (49 %, contre 46 %). La part du patrimoine bâti des communes exposée aux risques d'inondation par débordement de cours d'eau (31 %) ou submersion marine (3 %) est également comparable à celle des ménages, champ pour lequel une mesure comparable est produite par le SDES (SDES et ONRN, 2024), s'établissant à 28 % et 3,5 % respectivement.

La méthode permet de s'intéresser spécifiquement à l'exposition à certains risques (***figures 4 et 5***). Par exemple, 15 % des bâtiments des collectivités territoriales sont à risque fort de retrait-gonflement des sols argileux, pour une valeur vénale de 61 Md€ (soit 14 % du patrimoine bâti) ; les bâtiments les plus exposés à ce risque et les plus chers sont concentrés en Île-de-France et autour de la Méditerranée. De même, 31 % des bâtiments se trouvent dans des zones à risque d'inondation par remontée de cours d'eau au sens de l'EAIP (162 Md€, soit 38 % du patrimoine bâti).

Les estimations de la valeur et de l'exposition aux risques naturels de chaque bien détenu par les communes et les EPCI sont enfin combinées et utilisées pour expliquer les différences de niveau des dépenses d'assurance, telles qu'enregistrées dans les données comptables des collectivités. Toutes choses égales par ailleurs, une commune dont l'intégralité des biens est à risque d'inondation par remontée des cours d'eau s'acquitte d'une prime plus élevée de 8 % par rapport à celles dont aucun bien n'est sujet à ce risque. La surprime est de 5 % en cas de risque de retrait-gonflement des sols argileux.

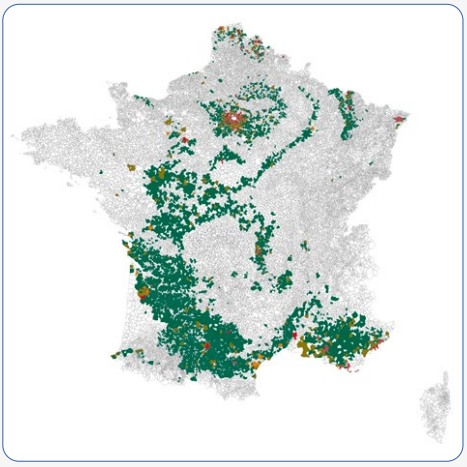
Figure 3 - Les données mobilisées pour l'analyse des dépenses d'assurance des collectivités



DRIAS : Donner accès aux scénarios climatiques Régionalisés français pour l'Impact et l'Adaptation de nos Sociétés et environnement.
SDES/ONERC : Service des données et études statistiques / Observatoire national sur les effets du réchauffement climatique.
DGPR : Direction générale de la prévision des risques.
BRGM : Bureau de recherches géologiques et minières.

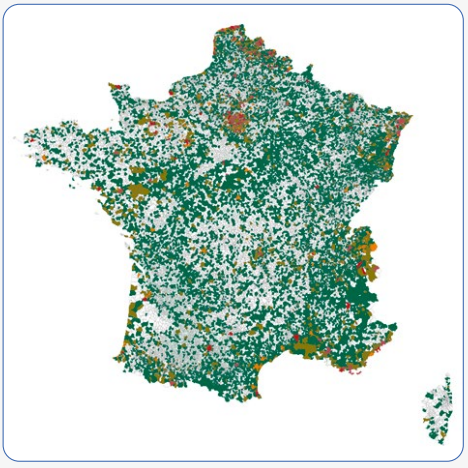
Les risques définis à une maille moins fine (séismes ou mouvements de terrain) ne sont pas significativement associés à des dépenses d'assurance plus élevées. Ces résultats ne valent que lorsque l'analyse est pondérée par la taille des communes : la relation positive entre exposition aux risques et dépenses d'assurance concerne les communes les plus peuplées. Ce résultat pourrait indiquer que les communes les plus peuplées et leurs assureurs connaissent davantage leur exposition au risque, et que ces derniers sont plus à même de la retranscrire dans le niveau des primes d'assurance.

▶ Figure 4 - Valeur vénale du patrimoine exposé au risque de retrait-gonflement des sols argileux



Source : DGFiP, demandes de valeurs foncières ; Bureau de recherches géologiques et minières (BRGM) ; calculs : inspection générale des finances (IGF) - pôle science des données.

▶ Figure 5 - Valeur vénale du patrimoine exposé au risque d'inondation par remontée de cours d'eau



Source : DGFiP, demandes de valeurs foncières ; Direction générale de la prévision des risques (DGPR) ; calculs : inspection générale des finances (IGF) - pôle science des données.

▶ La contribution de la science des données à l'IGF : une demande en forte croissance

Un levier de transformation, présentant néanmoins des limites

> ***Des équipes comme le pôle science des données de l'IGF répondent à une demande croissante d'analyses quantitatives, dans des délais souvent contraints, et contribuent à la solidité des constats établis par les missions de conseil au décideur.***

La demande en analyse de données est forte de la part des décideurs publics. Elle l'est en conséquence aussi pour les inspections. Grâce à leur flexibilité, des équipes comme le pôle science des données de l'IGF répondent à une demande croissante d'analyses quantitatives, dans des délais souvent contraints, et contribuent à la solidité des constats établis par les missions de conseil au décideur.

Ces équipes jouent un rôle d'agrégateur de la connaissance et des données produites par le système statistique public. La quantification existait auparavant à l'IGF, mais les missions devaient se débrouiller avec les données, en extrapolant des chiffrages existants ou en mobilisant elles-mêmes des méthodes qu'elles maîtrisaient. Les *data scientists*, à travers l'agrégation et la mise en perspective des données, apportent une dimension quantitative et systématique qui complète les approches généralement plus qualitatives des inspecteurs. Cela conduit en pratique à une forte participation des *data scientists* aux missions de conseil de l'inspection. Une trentaine de missions ont ainsi été accompagnées par le pôle science des données de l'IGF en 2023, dont quelques-unes portées uniquement par le pôle, soit une mission sur deux (IGF, 2024).

Par ailleurs, ce développement contribue fortement à l'acculturation des décideurs et des inspecteurs aux méthodes d'analyse quantitative, en se plaçant en intermédiaire, pour ainsi dire en pont, entre la statistique publique et les décideurs.

Ce mode de fonctionnement présente néanmoins certaines limites : s'il est utile de pouvoir mobiliser rapidement des données administratives ou des productions du système statistique public, il ne faut toutefois pas imaginer qu'il est possible de réaliser n'importe quelle étude sur cette base. Certaines études se heurtent à l'absence de données adaptées, les bases disponibles étant parfois trop anciennes, ne couvrant pas le périmètre nécessaire ou requérant un temps d'acculturation et de retraitement incompatible avec les délais de la mission. Il est également fréquent, mais c'est un problème très général en évaluation des politiques publiques, qu'une politique dans sa définition ne permette pas d'établir un contrefactuel, c'est-à-dire de trouver un point de comparaison pour apprécier l'impact de la mesure. Par exemple, si l'on cherche à évaluer un plan de soutien à un secteur entier de l'industrie, il est délicat de discerner un effet sur les entreprises traitées. Il faut alors recourir à des techniques plus élaborées comme le contrôle synthétique (Abadie et al., 2010), dont la mise en œuvre est susceptible de ne pas tenir dans les délais impartis, ou bien qui risque in fine de manquer de puissance statistique.

Les conditions d'une intégration réussie de la donnée dans la décision publique

L'expérience acquise dans le déploiement de la science des données au sein de l'IGF permet de dégager plusieurs enseignements pour maximiser son impact et favoriser son adoption.

Tout d'abord, il est très important de toujours faire preuve d'une grande pédagogie auprès des décideurs et des inspecteurs pour accompagner à chaque étape la remise des conclusions des analyses quantitatives. En effet, leur portée, leurs limites et leurs implications ne sont pas toujours évidentes pour les non-spécialistes. Bien expliquer les hypothèses sous-jacentes, les choix méthodologiques et les résultats est essentiel pour garantir une adoption éclairée et éviter tout malentendu. L'acculturation va d'ailleurs dans les deux sens : les *data scientists* eux-mêmes bénéficient des retours des inspecteurs et décideurs, qui apportent un regard critique et des connaissances opérationnelles indispensables. Cette interaction favorise une meilleure intégration des résultats dans les recommandations, en les rendant à la fois plus accessibles et plus actionnables.

> ***La fluidité dans les échanges de données et de documentation entre administrations apparaît comme un levier crucial.***

Ensuite, il est fondamental de pouvoir s'appuyer sur des infrastructures communes, sécurisées et robustes pour l'accès aux données. La diversité des sources mobilisées et la complexité des systèmes administratifs peuvent rendre ce travail particulièrement fastidieux et chronophage. Des infrastructures partagées comme le CASD permettent de standardiser les accès, garantir la sécurité et réduire le temps consacré à la collecte et à la préparation des données, laissant ainsi davantage de place à l'analyse et à la réflexion stratégique. Ces gains d'efficacité sont particulièrement précieux dans le cadre de missions souvent soumises à des délais contraints : il s'agit ici de lever le plus de freins possibles à la mobilisation de la donnée pour la décision publique.

Enfin, toujours dans cet objectif, la fluidité dans les échanges de données et de documentation entre administrations apparaît comme un levier crucial. Les travaux de science des données mobilisent souvent de multiples sources administratives. Or, la cartographie des bases de données et leur documentation sont parfois inexistantes, si bien que leur accès peut se révéler complexe. Harmoniser ces échanges, que ce soit par des protocoles simplifiés ou par des plateformes communes, permettrait non seulement de fluidifier les travaux, mais aussi de mieux exploiter le potentiel des données publiques, comme l'exemple du CASD le montre avec brio.

Le développement de la démarche

Le développement d'un pôle dédié à la science des données a été permis par l'ouverture et la mise à disposition des données publiques. Dans le contexte récent de réinternalisation des missions des cabinets privés au sein de l'État, l'IGF est ainsi en mesure de répondre à un panel plus large de missions, en mobilisant à plein les capacités de ses *data scientists*.

Comme évoqué, d'autres inspections ou instances d'évaluation suivent une démarche similaire afin d'appuyer ou d'approfondir leurs constats à l'aide des données. C'est notamment le cas de l'inspection générale des affaires sociales (Igas)[19], de l'inspection générale de l'environnement et du développement durable (IGEDD) et de l'inspection générale de l'éducation, du sport et de la recherche (IGESR), mais aussi de la Cour des comptes, preuve de l'intérêt suscité au sein de ces services par la perspective de conduire en interne une partie des analyses quantitatives dont les missions ont besoin.

Ces développements s'inscrivent dans une démarche plus large de renforcement de l'écosystème public de la science des données et de l'évaluation des politiques publiques, incluant notamment le monde académique (via des institutions comme l'Institut des politiques publiques). Cette démarche renforce par là même la sensibilisation des décideurs aux enjeux de l'évaluation et du recours aux données pour réaliser leurs arbitrages.

19 Voir l'article de Juliette Berthe dans ce même numéro.

▶ Fondements juridiques

- Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique. In : *site de Légifrance*. [en ligne]. [Consulté le 18 décembre 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746*.

▶ Bibliographie

- ABADIE, Alberto, DIAMOND Alexis et HAINMUELLER, Jens, 2010. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. In : *Journal of the American Statistical Association*. [en ligne]. Juin 2010, Vol. 105, N° 490, pp. 493-505. [Consulté le 18 décembre 2024]. Disponible à l'adresse : *https://www.tandfonline.com/doi/abs/10.1198/jasa.2009.ap08746*.

- ANDRÉ, Mathias et MESLIN, Olivier, 2022. Patrimoine immobilier des ménages : enseignements d'une exploitation de sources administratives exhaustives. In : *Courrier des statistiques*. [en ligne]. 20 janvier 2022. Insee. N° N7, pp. 107-125. [Consulté le 10 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6035944?sommaire=6035950*.

- ANDRÉ, Mathias et MESLIN, Olivier, 2025. Le bonheur est dans le prix : estimation du patrimoine immobilier brut des ménages sur données administratives exhaustives. In : *Documents de travail*. [en ligne]. 10 février 2025. Insee. N° 2025-04. [Consulté le 14 février 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/8351238*.

- BLANCHET, Didier, 2020. Des modèles de microsimulation dans un institut de statistique - Pourquoi, comment, jusqu'où ? In : *Courrier des statistiques*. [en ligne]. 29 juin 2020. Insee. N° N4, pp. 6-22. [Consulté le 18 décembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4497054?sommaire=4497095*.

- BOTHOREL, Éric, 2020. Pour une politique de la donnée. In : *Rapport de la mission sur la politique publique de la donnée, des algorithmes et des codes sources*. [en ligne]. 23 décembre 2020. [Consulté le 18 décembre 2024]. Disponible à l'adresse : *https://www.info.gouv.fr/rapport/11979-rapport-sur-la-politique-publique-de-la-donnee-des-algorithmes-et-des-codes-sources*.

- CHEN, Tianqi et GUESTRIN, Carlos, 2016. XGBoost: A Scalable Tree Boosting System. In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [en ligne]. 13 août 2016. Association for Computing Machinery. Pp. 785-794. [Consulté le 18 décembre 2024]. Disponible à l'adresse : *https://dl.acm.org/doi/10.1145/2939672.2939785*.

- COMTE, Frédéric, DEGORRE, Arnaud et LESUR, Romain, 2022. Le SSPCloud : une fabrique créative pour accompagner les expérimentations des statisticiens publics. In : *Courrier des statistiques*. [en ligne]. 20 janvier 2022. Insee. N° N7, pp. 68-87. [Consulté le 10 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6035940?sommaire=6035950*.

- DE PERETTI, Gaël et TOUCHELAY, Béatrice, 2024. Statistiques publiques et débat démocratique : de nouvelles attentes et de nouveaux enjeux (1988-2016). In : *Courrier des statistiques*. [en ligne]. 8 juillet 2024. Insee. N° N11, pp 11-30. [Consulté le 10 mars 2025. Disponible à l'adresse : *https://www.insee.fr/fr/information/8203034?sommaire=8203072*.

- FREUND, Yoav et SCHAPIRE, Robert E., 1996. *Experiments with a New Boosting Algorithm*. In : *site de AT&T*. [en ligne]. 22 janvier 1996. [Consulté le 18 décembre 2024]. Disponible à l'adresse : *https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d186abec952c4348870a73640bf849af9727f5a4*.

- GADOUCHE, Kamel, 2019. Le Centre d'accès sécurisé aux données (CASD), un service pour la *data science* et la recherche scientifique. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 76-92. [Consulté le 18 décembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254227?sommaire=4254170*.

- GIVORD, Pauline, 2014. Méthodes économétriques pour l'évaluation de politiques publiques. In : *Économie et Prévision*. [en ligne]. Direction générale du Trésor. N° 204-205, pp. 1-28. [Consulté le 18 décembre 2024]. Disponible à l'adresse : *https://doi.org/10.3917/ecop.204.0002*.

- HELFENSTEIN Xavier, 2022. La base permanente des équipements (BPE). Une source statistique singulière et constamment en mouvement. In : *Courrier des statistiques*. [en ligne]. 29 novembre 2022. Insee. N° N8, pp. 131-148. [Consulté le 18 décembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665194?sommaire=6665196*.

- INSPECTION GÉNÉRALE DES FINANCES, 2024. *Rapport d'activité de l'année 2023*. [en ligne]. Mai 2024. [Consulté le 10 mars 2025]. Disponible à l'adresse : *https://www.igf.finances.gouv.fr/sites/igf/accueil/nos-activites/nos-rapports-dactivite.html*.

- KOUMARIANOS, Heïdi, LEFEBVRE, Olivier et MALHERBE, Lucas, 2024. Les appariements : finalités, pratiques et enjeux de qualité. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2024. Insee. N° N11, pp 117-139. [Consulté le 18 décembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8203044?sommaire=8203072*.

- MARTIN, Olivier, 2023. *Chiffre*. Anamosa, collection le mot est faible. ISBN 978-2-38191-054-3. 11 mars 2023.

- PRADA, Michel 2012. Le métier d'inspecteur : de la vérification à l'évaluation, 1946-2009. In : *Dictionnaire historique des inspecteurs des Finances 1801-2009*. [en ligne]. IGPDE. Pp. 121-126. [Consulté le 10 mars 2025]. Disponible à l'adresse : *https://books.openedition.org/igpde/3623*.

- RIVIÈRE, Pascal, 2020. Qu'est-ce qu'une donnée ? – Impact des données externes sur la statistique publique. In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. Insee. N° N5, pp. 114-131. [Consulté le 18 décembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5008707?sommaire=5008710*.

- SERVICE DES DONNÉES ET DES ÉTUDES STATISTIQUES et OBSERVATOIRE NATIONAL DES RISQUES NATURELS, 2024. *Chiffres clés des risques naturels - Édition 2023*. [en ligne]. 30 janvier 2024. [Consulté le 18 décembre 2024]. Disponible à l'adresse : *https://www.statistiques.developpement-durable.gouv.fr/chiffres-cles-des-risques-naturels-edition-2023*.

- VOLLE, Michel, 1984. *Le métier de statisticien*. Economica. ISBN 2-7178-0824-8.

Le défi des données pour l'inspection générale des affaires sociales

Juliette Berthe *

L'inspection générale des affaires sociales (Igas) mène des missions d'audit et d'évaluation dans les domaines de la santé, du travail et des solidarités. Elle s'est dotée d'un pôle *data* pour asseoir les travaux de l'inspection sur des éléments quantifiés et objectiver en particulier certains points ou recommandations. Les *data scientists* doivent exploiter des données sur des thématiques variées, répondre aux besoins des missions dans des délais courts et composer à la fois avec des bases bien structurées, notamment dans le champ de la santé et de l'emploi, et des systèmes d'information locaux hétérogènes et fragmentés, notamment dans le domaine des solidarités.

Le travail du *data scientist* est vaste : il doit rechercher, croiser et nettoyer des données qui peuvent être incomplètes ou dispersées, en garantissant leur pertinence et leur interprétation correcte. Lorsqu'elles sont absentes, il recourt à des approches comme le *web scraping*. Il doit aussi arbitrer entre fraîcheur et fiabilité des données, privilégiant parfois des sources officielles mais datées pour assurer la robustesse des analyses.

Son rôle ne se limite pas à la *data science* : il interagit avec les inspecteurs et les acteurs de terrain pour contextualiser ses analyses. En combinant rigueur, adaptation et pragmatisme, il éclaire les décisions publiques malgré des contraintes méthodologiques et temporelles fortes, dans l'objectif de contribuer à des politiques plus informées et efficaces.

The General Inspectorate of Social Affairs (IGAS) conducts audit and evaluation missions in the fields of health, labor, and social solidarity. It has set up a data pole to consolidate the inspectorate's projects by using quantified elements and particularly to substantiate some points or recommendations. Data scientists have to process thematically diverse data, meet the mission needs within short timeframes, and work with well-structured databases, in the fields of health and employment notably, as well as with heterogeneous and fragmented local information systems, as is particularly the case with social solidarity.

The data scientist has a wide scope of activities: they have to search for, cross-reference, and clean data that is sometimes incomplete or scattered, ensuring its relevance and accurate interpretation. When data is missing, they turn to alternative approaches such as web scraping. They also have to balance between data freshness and reliability, sometimes favouring official but older sources to ensure the robustness of analyses.

Their role extends beyond data science: they collaborate with inspectors and field actors to provide context to their analyses. By combining rigor, adaptability, and pragmatism, they help inform public decision-making despite strong methodological and time constraints, ultimately aiming to contribute to more effective and well-informed policies.

* Responsable du pôle *data*, inspection générale des affaires sociales (Igas).
juliette.berthe@igas.gouv.fr

Pour orienter leurs choix stratégiques, les décideurs publics ont besoin d'analyses sérieuses et rigoureuses. Que ce soit pour concevoir des politiques publiques, estimer l'impact d'une mesure, élaborer de nouvelles lois ou allouer des ressources, ils doivent s'appuyer sur des études approfondies et impartiales menées par des spécialistes du sujet. Les inspections ministérielles apportent cette expertise précieuse, chacune dans son domaine spécifique, ou en collaboration lorsque la complexité du sujet exige une approche multidisciplinaire. Dans le domaine social, cette responsabilité revient à l'inspection générale des affaires sociales[1] (Igas). Son périmètre d'intervention couvre des enjeux majeurs mobilisant une part significative des ressources nationales et affectant directement la vie de tous les citoyens : emploi, travail et formation professionnelle, santé publique, organisation des soins, cohésion sociale, sécurité sociale, protection des populations.

La disponibilité croissante de larges bases de données administratives ou d'enquêtes (d'accès public ou sur demande motivée), conjuguée à un contexte d'innovation en matière de méthodes quantitatives, constitue une véritable opportunité pour l'enrichissement et la pertinence des constats et recommandations des missions confiées à l'Igas. L'inspection a donc créé en 2023 un pôle *data* où ces aspects d'analyse sont pris en charge par des *data scientists*[2] dédiés. Il comprend aujourd'hui trois *data scientists* permanents et un étudiant en apprentissage. Il a déjà appuyé vingt missions réparties équitablement sur l'ensemble des champs que recouvre l'inspection.

> ***Si ces experts de la science des données doivent développer des indicateurs pertinents, dans un esprit de rigueur, de transparence et d'objectivité, leur approche diffère sensiblement des pratiques de la statistique publique sur de nombreux aspects.***

Si ces experts de la science des données doivent développer des indicateurs pertinents, dans un esprit de rigueur, de transparence et d'objectivité, leur approche diffère sensiblement des pratiques de la statistique publique sur de nombreux aspects. En effet, plusieurs spécificités encadrent strictement le travail du *data scientist* : l'ampleur du champ d'action de l'Igas, la nature et la source des données disponibles et, surtout, l'objectif des missions, qui est d'éclairer le décideur, souvent dans des délais très courts, sur des problématiques généralement très ciblées. Cet objectif impose des contraintes temporelles et une utilisation précise des données. Pourtant, c'est justement ce cadre exigeant, à la fois stimulant et complexe, qui constitue la spécificité et l'intérêt de ce travail et son caractère spécial.

▶ Une culture de la donnée très variable selon les champs —

L'Igas opère dans trois domaines : le travail, la santé et les solidarités. La richesse des ressources en données et leur niveau de maturité varient grandement selon ces champs. Les données portant sur le travail et la santé sont les plus robustes.

1 *https://igas.gouv.fr/.*
2 Voir la définition du métier de *data scientist* dans Bourlange et al. (2021).

La santé, très acculturée aux données

Dans le secteur de la santé, la culture de la donnée est ancrée depuis longtemps chez la plupart des acteurs. À l'hôpital, la saisie de données à des fins à la fois financières et épidémiologiques a débuté dans les années quatre-vingts. Au fur et à mesure, ces saisies ont porté sur davantage d'informations ou ont été enrichies d'autres données. Aujourd'hui, l'intérêt de faire remonter de l'information et de la consolider est partagé par l'ensemble des acteurs et le processus fait partie de leur quotidien. Par ailleurs, les données administratives sont particulièrement riches grâce au **système national des données de santé**[3] (SNDS) (***encadré 1***). Ce dernier contient notamment des informations très détaillées sur la consommation de soins à la maille du patient, que ce soit des soins en ville ou en établissement de santé. Il s'agit principalement du détail des remboursements des dépenses par l'assurance maladie et des données décrivant l'ensemble des séjours à l'hôpital. Par ricochet, ce niveau de détail renseigne très précisément sur l'activité des professionnels et des établissements. Les variables financières des établissements publics sont également communiquées et consolidées à une échelle nationale.

> ***Les données contenues dans le SNDS sont structurées et standardisées. Elles obéissent à des nomenclatures établies et partagées nationalement, voire internationalement.***

Les données contenues dans le SNDS sont structurées et standardisées. Elles obéissent à des nomenclatures établies et partagées nationalement, voire internationalement pour certaines. Par exemple, les causes de décès sont codées selon la classification internationale des maladies de l'Organisation mondiale de la santé (Coudin et Robert, 2024). Grâce à cette standardisation, l'interopérabilité interne et externe du système est garantie.

En complément de ces données individuelles, les agences régionales de santé (ARS) utilisent divers outils de gestion pour piloter l'offre de soins, gérer les financements et suivre les dépenses, notamment ceux des établissements. Enfin, les enquêtes et études menées en particulier par la direction de la recherche, des études, de l'évaluation et des statistiques (Drees) apportent des données particulièrement précieuses sur une multitude de sujets qui vont au-delà du champ de la santé à strictement parler ; le champ médico-social y est notamment abordé. Une grande part de ces données est disponible en open data sur son site. Les données plus confidentielles, notamment celles relevant d'enquêtes spécifiques, sont pour la plupart disponibles au Centre d'accès sécurisé aux données (CASD) [Gadouche, 2019].

Le champ travail possède également des données riches et structurées

La culture de la donnée est également très présente dans le domaine du travail au sens large ; on parle alors du champ TEFP (travail, emploi et formation professionnelle). La principale source de données administratives est la **déclaration sociale nominative** (DSN) : elle apporte des informations individuelles sur les salariés, leurs contrats de travail, les rémunérations et primes reçues, les cotisations versées, les absences et

3 *https://www.snds.gouv.fr/SNDS/Accueil.*

▶ Encadré 1. Le SNDS retrace le parcours de soins de chaque assuré social

En France, le système national des données de santé (SNDS) rassemble les principales données de santé provenant de sources administratives et médicales. Géré par la plateforme des données de santé (PDS ou *Health data hub**) et la caisse nationale de l'assurance maladie (Cnam), il est issu de la mise en relation du système national d'information interrégimes de l'assurance maladie (Sniiram), du programme de médicalisation des systèmes d'information (PMSI) et de la base des causes médicales de décès (BCMD).

Le Sniiram, géré par la Cnam, contient les données relatives à toutes les dépenses d'assurance maladie. Il est composé d'une base de données individuelles sur la consommation de soins, appelée *datamart*** de consommation interrégimes (DCIR), de 15 bases thématiques de données agrégées et d'un échantillon au 2/100e de patients ayant bénéficié d'un soin, destiné à des études longitudinales.

Le PMSI, géré par l'Agence technique de l'information sur l'hospitalisation (ATIH), a été créé en 1982 pour permettre l'analyse de l'activité médicale des établissements de santé à des fins d'allocation budgétaire. Il intègre des informations administratives et médicales relatives à chaque séjour dans un établissement de santé, public ou privé, pour tout type d'hospitalisation : médecine, chirurgie et obstétrique (MCO), soins médicaux et de réadaptation (SMR), hospitalisation à domicile (HAD) et psychiatrie.

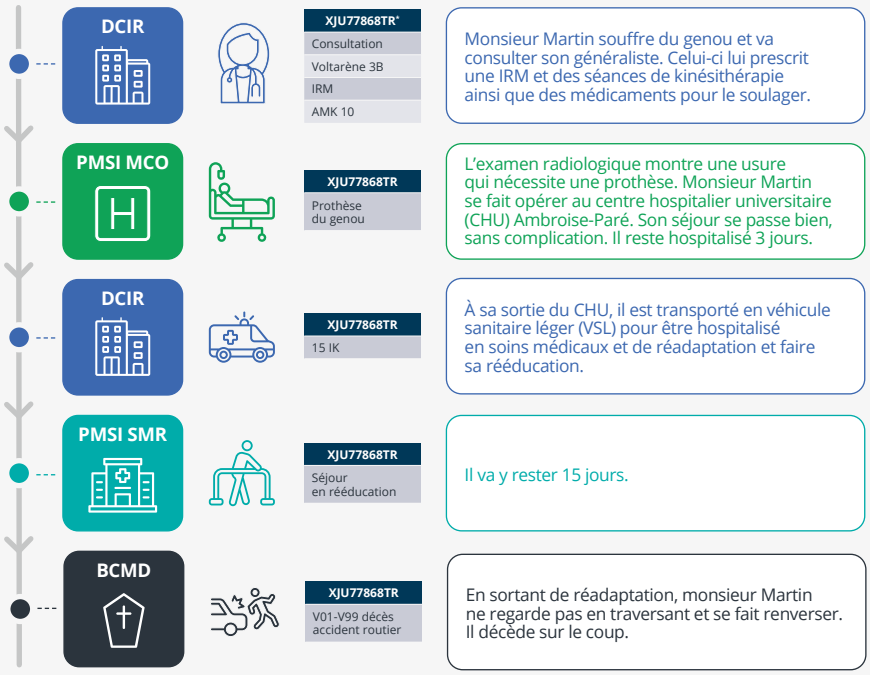
Enfin, la BCMD, gérée par le Centre d'épidémiologie sur les causes médicales de décès (CépiDC), contient les données relatives aux causes de décès pour chaque individu.

En résumé, le SNDS fournit des informations détaillées :
- sur l'ensemble des dépenses liées aux soins, qu'ils soient réalisés en ville ou en établissement de santé et quel que soit le type de dépense : consultation médicale ou paramédicale, dépense de pharmacie, transport sanitaire, etc. ;
- sur l'ensemble des séjours en établissement : durée, passage aux urgences, coût, mais aussi diagnostics et comorbidités du patient.

Les trois systèmes Sniiram, PMSI et BCMD permettent, par le biais d'un identifiant unique, de chaîner les données à la granularité du patient. On obtient ainsi une vision globale du parcours de soins des assurés sociaux jusqu'à la cause de leur décès (***figure encadré***).

* *https://www.health-data-hub.fr/.*
** *Magasin de données.*

Le parcours de soins de monsieur Martin dans le SNDS



* XJU77868TR est l'identifiant (fictif) de monsieur Martin dans le SNDS. C'est cet identifiant unique dans l'ensemble des bases du SNDS qui permet de suivre son parcours de soin.

reprises, etc. (Humbert-Bottin, 2018). Très normalisée, mais nécessitant beaucoup de retraitements pour la production de statistiques, elle est exploitée à cette fin par une multitude d'acteurs, notamment par l'Insee, qui produit à partir de celle-ci la base Tous salariés[4] (Brunet et al., 2023). Si l'Igas n'exploite pas encore directement la DSN, elle s'appuie très souvent sur la base Tous salariés.

Ces données sur l'emploi des salariés issues des processus administratifs sont complétées par d'autres données de gestion sur l'emploi non salarié, d'une part, et sur le chômage et la formation, d'autre part, provenant notamment de la direction générale à l'emploi et à la formation professionnelle (DGEFP), de France Travail et de France compétences. Concernant la formation professionnelle, la Caisse des dépôts a par ailleurs développé une plateforme de mutualisation et d'échange de données, AGORA[5]. Son objectif est de consolider en temps réel les informations sur les parcours et d'améliorer ainsi le pilotage dans le domaine. Elle est alimentée par l'ensemble des acteurs impliqués : organismes formateurs, financeurs, rémunérateurs et certificateurs. Enfin, des études et enquêtes menées notamment par l'Insee et la direction de l'animation de la recherche, des études et des statistiques (Dares) complètent la vision du champ. La plupart de ces données sur le champ TEFP sont accessibles par l'Igas, soit en open data, soit par le CASD après un accord des producteurs.

Les solidarités : un champ plus diversifié et moins centralisé

Le champ des solidarités est particulièrement diversifié ; il regroupe des acteurs et des activités très variés. De manière générale, il concerne le soutien aux personnes vulnérables en raison de leur âge, de leur handicap ou de leur situation familiale ou sociale complexe. Les politiques concernées sont notamment celles de l'accueil de la petite enfance, de la protection maternelle et infantile, de la protection de l'enfance, du soutien aux personnes handicapées ou en perte d'autonomie ainsi qu'à leurs proches aidants, de la lutte contre la pauvreté, de l'aide aux familles nombreuses et aux familles monoparentales. Les interventions relèvent à la fois du soutien financier, de l'accompagnement, de l'aide à domicile, de l'accueil en institution, voire de la prévention et de l'éducation.

> ***Sur le volet médico-social, la mise à disposition et l'exploitation des données sont plus difficiles.***

La diversité des publics et des actions menées conduit à mobiliser un large éventail d'acteurs, à tous les niveaux de l'échelle territoriale. Au niveau local, les communes et les départements sont chefs de file de l'action sociale : ils jouent un rôle clé dans le financement et la gestion des services sociaux et médico-sociaux. En région, les ARS se concentrent sur la qualité et la régulation de l'offre médico-sociale liée à la santé, tandis que les directions régionales de l'économie, de l'emploi, du travail et des solidarités (Dreets) œuvrent sur la dimension sociale et professionnelle : elles coordonnent les politiques sociales et travaillent notamment sur l'inclusion sociale et professionnelle. Enfin, au niveau national, la Caisse nationale de solidarité pour

4 *https://www.insee.fr/fr/metadonnees/source/serie/s1998.*
5 *https://travail-emploi.gouv.fr/agora-la-plateforme-dechange-de-donnees-de-la-formation-professionnelle.*

l'autonomie (CNSA) coordonne les politiques publiques de soutien à l'autonomie sur les aspects de financement, pilotage des maisons départementales des personnes handicapées (MDPH)[6], planification et suivi des politiques médico-sociales. La caisse nationale des allocations familiale (Cnaf) et la mutualité sociale agricole (MSA) gèrent quant à elles les prestations sociales à travers leurs réseaux de caisses.

Pour étudier le champ des solidarités, l'Igas a à sa disposition les enquêtes et publications de la Drees et les données de gestion des collectivités et des caisses. On comprend aisément que l'organisation complexe du champ, à la fois par la multiplicité des domaines d'intervention et par l'autonomie plus ou moins grande laissée aux acteurs locaux dans la gestion, conditionne fortement la disponibilité des données, leur homogénéité et leur interopérabilité. D'une manière générale, on observe que l'existence d'un réseau dédié piloté permet des remontées harmonisées. Ainsi, les prestations familiales et de solidarité qui sont versées par les réseaux de la Cnaf et de la MSA sont agrégées dans une base nationale à laquelle l'Igas a accès via le CASD. En revanche, sur le volet médico-social, la mise à disposition et l'exploitation des données sont plus difficiles. En effet, les canaux pour saisir et remonter l'information sont multiples et ne sont pas tous harmonisés. La diversité des acteurs (MDPH, ARS, conseils départementaux, centres communaux d'action sociale) et le fait qu'ils ne soient pas entièrement dédiés à la problématique médico-sociale complexifient le sujet. Sur le champ du handicap, la CNSA travaille à moderniser la gestion des données médico-sociales, notamment par la création d'un système d'information harmonisé et la mise en place d'un centre de données. Elle vise en particulier à construire un système d'information sur les MDPH, permettant d'homogénéiser les informations autour de l'accompagnement par ces services des personnes handicapées, afin d'en donner une vision nationale. Cette base de données contiendra les informations sur les différentes demandes d'accompagnement formulées pour les personnes handicapées et le déroulement de leur instruction. Un appariement avec le SNDS est par ailleurs prévu pour pouvoir observer les liens entre ces accompagnements et le parcours de santé des bénéficiaires.

Plus largement, même si la Drees s'attache à centraliser des informations locales via des enquêtes ou des remontées de données individuelles[7], il subsiste que les données relatives au champ des solidarités ne bénéficient pas d'une normalisation comme celles de la santé ou du travail. Leur analyse est donc bien plus complexe (Cotton et Haag, 2023) et la robustesse des résultats moins assurée.

6 Les MDPH sont, dans chaque département, le guichet unique d'accès simplifié aux droits et prestations pour les personnes handicapées.

7 Par exemple via l'enquête Aide sociale réalisée auprès des départements (Diallo et al., 2024) ou les remontées de données individuelles sur l'orientation, l'accompagnement et l'insertion des bénéficiaires du revenu de solidarité active (RSA).

▶ Les spécificités du travail des *data scientists* à l'Igas

Intervenir dans un contexte de missions, sur des sujets très divers, des problématiques très précises et dans une fenêtre de temps limitée, conditionne fortement le travail des *data scientists* de l'Igas.

Des délais contraints qui obligent aux compromis

Le rôle du pôle *data* est d'apporter un éclairage spécifique à certaines missions de l'Igas au travers d'analyses quantitatives. Il peut s'agir d'une statistique qui permettra d'appuyer une recommandation, de détection d'anomalies pour mieux cibler des contrôles, d'une classification pour dégager des comportements ou situations similaires, etc. Chaque approche *data* répond à un besoin précis pour la mission.

Les données sur un sujet sont parfois riches et abondantes. Pour autant elles ne répondent pas nécessairement aux nécessités des missions de l'Igas. Leur fraîcheur est notamment un problème récurrent. L'Igas s'appuie beaucoup sur des données officielles, issues d'enquêtes ou de traitements de données administratives réalisés par la statistique publique (Insee ou services statistiques ministériels, SSM). La qualité des données est alors indiscutable, mais la contrepartie réside dans le délai d'obtention. Le risque est que l'Igas dresse des constats ou appuie ses recommandations sur des informations qui pourraient ne plus refléter, ou que partiellement, la réalité du moment. Il est donc nécessaire que le *data scientist* apprécie avec les inspecteurs de la mission les enjeux de temporalité.

> ***Les données sur un sujet sont parfois riches et abondantes. Pour autant elles ne répondent pas nécessairement aux nécessités des missions de l'Igas.***

Ainsi, dans le cadre d'une mission réalisée en 2024 visant à évaluer le caractère contraint des temps partiels dans certains secteurs d'activité, l'Igas s'est appuyée sur la base Tous salariés de l'Insee. L'objectif était de mesurer la part des salariés travaillant à temps partiel dans certains secteurs choisis, mais aussi le nombre d'emplois cumulés par ailleurs par chacun de ces salariés, pour mieux appréhender le caractère contraint de leur temps partiel (Magnier et Viossat, 2024). À l'époque où les travaux ont été réalisés, le millésime le plus récent de la base Tous salariés était 2022, pour un rapport publié fin 2024 : les constats portaient donc sur des données datant de deux ans. Il aurait été possible de travailler sur des informations plus actuelles issues directement de la DSN. Cependant, les analyses auraient alors reposé sur des données administratives brutes, donc de qualité moindre, notamment sur le nombre d'heures travaillées ; ceci aurait alors fragilisé les constats. L'Igas a finalement choisi de privilégier la fiabilité de l'information, quitte à s'appuyer sur des données moins récentes.

Savoir trouver les données disponibles et les mettre en musique

Une partie importante du travail du *data scientist* consiste à explorer la multitude de bases de données disponibles, qu'il s'agisse de ressources accessibles en open data ou via des dispositifs sécurisés comme le CASD. Il est en effet crucial de trouver la bonne source d'information. La mission « Lieux de vie et accompagnement des personnes âgées en perte d'autonomie » (Emmanuelli et al., 2023) est un exemple parlant sur le potentiel d'analyse que représentent les données en accès libre. En effet, en mobilisant les informations diffusées sur les sites de la Drees et de l'Insee, il a été possible de caractériser à l'échelle départementale la demande et l'offre de prise en charge de personnes dépendantes :

- d'un côté, les données en accès libre ont permis de décrire comment les personnes âgées dépendantes se répartissent dans les départements, quelle part elles représentent dans la population totale de chaque département et comment la situation est susceptible d'évoluer à l'horizon 2030 ou 2040 ;
- d'un autre côté, les départements ont pu être classés selon quatre catégories d'offre de prise en charge, à partir de sept indicateurs sur l'offre en ville et en établissement.

La confrontation de ces deux analyses permet ainsi à chaque département de se situer du point de vue de l'offre et de la demande (***figure***) : elle constitue un outil concret pour se préparer au défi démographique qui arrive. Par ailleurs, comme ce travail s'appuie sur des données accessibles à tous, les acteurs locaux sont en capacité de le reproduire et de l'actualiser facilement. Ils peuvent également le compléter à partir d'informations qui leur sont propres ou qu'ils jugeraient pertinentes.

Savoir exploiter les données brutes locales : en investissant dans la connaissance du terrain…

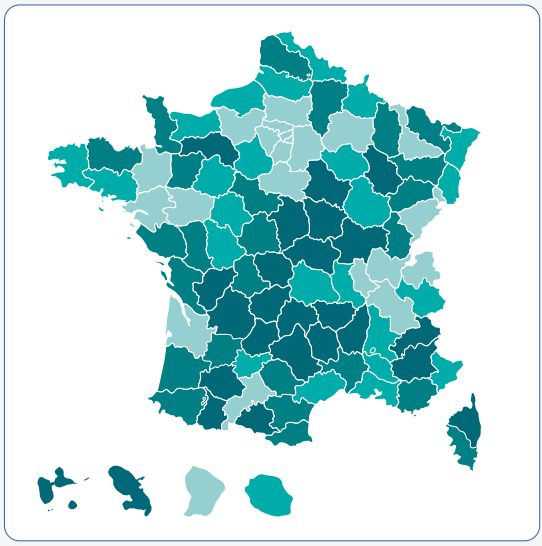
S'il est confortable de travailler sur des données structurées et bien normées, comme celles évoquées précédemment, elles ne suffisent pas toujours à répondre aux besoins diversifiés de l'Igas. Les données des systèmes d'information (SI) locaux sont alors une source précieuse d'information. Cependant, elles nécessitent un véritable investissement en vue de leur utilisation. En effet, il s'agit de données brutes de production qui n'ont pas été élaborées à des fins statistiques. Le *data scientist* doit donc réaliser un important travail pour en comprendre le sens, les adapter sur le plan sémantique et réaliser des ajustements techniques, ceci afin d'homogénéiser la structure des fichiers et des variables. Par ailleurs, les SI étant alimentés en permanence, il est essentiel de définir la temporalité sur laquelle va porter l'analyse et de figer les données à un instant t. Les situations individuelles qui y figurent ne sont alors pas nécessairement à jour : il faut tenir compte de cet aspect dans l'analyse.

« S'il est confortable de travailler sur des données structurées et bien normées, elles ne suffisent pas toujours à répondre aux besoins diversifiés de l'Igas. Les données des systèmes d'information (SI) locaux sont alors une source précieuse d'information qui nécessite un véritable investissement en vue de leur utilisation. »

▶ Figure - Le potentiel d'analyse des données en accès libre : illustration

Dans le cadre de la mission « Lieux de vie et accompagnement des personnes âgées en perte d'autonomie » (Emmanuelli et al., 2023), il était important d'éclairer la situation des départements en matière d'offre et de demande de prise en charge de personnes âgées dépendantes. Les données en accès libre de la Drees et de l'Insee permettent d'éclairer les besoins à venir et les forces et faiblesses de l'offre actuelle.

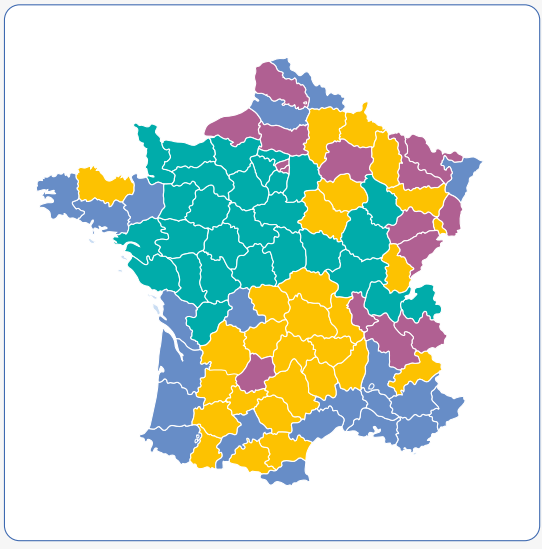
a Part des personnes âgées en perte d'autonomie dans la population en 2040



Champ : France hors Mayotte.
Source : Données issues du modèle LIVIA de la Drees 2023.

b Caractéristiques de l'offre de prise en charge des personnes âgées en perte d'autonomie

Selon sept variables caractéristiques de l'offre de prise en charge des personnes âgées en perte d'autonomie, une typologie en quatre classes a été construite à partir d'une méthode statistique de regroupement de départements proches.



Champ : France métropolitaine.
Source : Données Drees – Insee et moissonnage d'annuaires en ligne de résidences services effectué par le pôle *data* de l'Igas.
Mise en forme : Igas.

Cette approche est relativement fréquente, notamment dans le champ médico-social où les données nationales ne sont pas toutes consolidées. Ainsi, dans les missions portant sur l'aide sociale à l'enfance (ASE), l'Igas n'a, pour certains départements, d'autre source que celle des SI des conseils départementaux[8] (***encadré 2***). L'investissement du *data scientist* pour comprendre les données contenues dans ces SI est essentiel : il passe par

▶ Encadré 2. La difficulté à exploiter les données extraites des SI locaux : l'exemple de l'aide sociale à l'enfance (ASE)

L'exemple du suivi par les départements de l'aide sociale à l'enfance (ASE) illustre les difficultés que peut rencontrer l'Igas pour exploiter à des fins statistiques les données des systèmes d'informations locaux.

Les départements n'utilisent pas tous les mêmes logiciels de gestion

Pour le suivi des jeunes de l'ASE, les départements ont une autonomie totale pour choisir leur logiciel de gestion, qui peut donc varier d'un département à l'autre. Ainsi, dans le cadre de sa dernière mission sur l'ASE, l'Igas a étudié la situation de quatre départements. Trois d'entre eux utilisent le logiciel IODAS, tandis qu'un autre a développé son propre outil, webASE. La précédente mission avait été réalisée dans un département qui utilisait SOLIS. Du fait de cette hétérogénéité, les modalités d'extraction des données, la nature des éléments recueillis et le format des données diffèrent d'un département à l'autre.

Les logiciels utilisés ne sont pas conçus pour l'extraction massive de données

Les logiciels utilisés pour le suivi des jeunes de l'ASE ne sont pas spécifiquement conçus pour l'extraction et le traitement de grandes quantités de données. Par exemple, IODAS et SOLIS reposent sur l'outil BusinessObject pour exécuter des requêtes, ce qui limite la fluidité et l'efficacité des extractions massives de données. Cette contrainte technique et l'hétérogénéité des logiciels se traduisent par une grande variabilité dans les formats et la structure des informations extraites, avec l'impossibilité d'automatiser d'un département à l'autre les processus d'extraction.

Les départements n'ont pas tous les mêmes pratiques de gestion des données

Les pratiques de gestion des données varient également selon les départements. Certains d'entre eux adoptent une approche « multidomaine » : un même outil, comme IODAS, est utilisé pour gérer non seulement les données relatives à l'ASE, mais aussi celles concernant d'autres secteurs tels que l'insertion sociale. Cette gestion multiforme des données complique considérablement les demandes d'extraction, qui doivent tenir compte de la diversité des services intégrés dans un même logiciel.

La qualité des données est également hétérogène et dépend des pratiques

Une fois les données extraites, leur qualité dépend fortement des pratiques de saisie humaine. Selon les préférences des agents en charge du suivi, de nombreux champs sont soit laissés libres, soit remplis selon des règles hétérogènes. Ainsi, des variations typographiques peuvent apparaître pour des informations identiques (par exemple « M. », « Mr » ou « Monsieur » pour une même civilité), ou des incohérences dans le format des dates (par exemple « 01/09/24 » ou « 2024-09-01 »), ce qui complique la standardisation et l'analyse. De plus, certaines erreurs de saisie, ou des conventions locales comme l'utilisation de la date « 2000-01-01 » pour signifier une valeur manquante, rendent difficile la distinction entre des données valides et des données erronées.

Des données essentielles peuvent manquer (ou être codées en « ne sait pas »)

Un autre problème majeur est l'absence fréquente d'éléments essentiels, concernant notamment les signalements reçus au sujet de l'enfant. Ces données ne sont pas systématiquement renseignées, ce qui empêche une évaluation complète et précise du vécu de l'enfant dans le système de protection. De même, la catégorie des « types de mesure » est souvent incomplète, avec de nombreuses valeurs codées en « ne sait pas », ce qui ajoute à la difficulté d'analyse.

Les changements de logiciel peuvent rendre complexe l'analyse des évolutions

Enfin, des changements de logiciels, comme celui intervenu en juin 2018 pour un département, peuvent introduire des différences significatives dans la structure des données entre les périodes précédant et suivant la migration. Dans l'exemple évoqué, certaines décisions n'ont pas été transférées lors de la migration : des appariements complexes ont dû être mis en œuvre entre les données anciennes et nouvelles pour assurer leur cohérence. L'introduction du dispositif Olinpe a par ailleurs ajouté de nouveaux champs, qui ne sont pas compatibles avec le format initial du logiciel, ce qui complique encore la gestion des données et leur analyse comparative.

8 La Drees constitue une base nationale consolidée sur les enfants de l'ASE dénommée Olinpe (Observation longitudinale, individuelle et nationale en protection de l'enfance). Ce dispositif n'est pas encore exhaustif : en 2023, 32 départements ont transmis leurs données.

des échanges constants avec les spécialistes du métier. Il peut également accompagner la mission dans ses déplacements sur le terrain, afin de mieux cerner ce que représentent les données, mais aussi le contexte dans lequel elles sont produites. Cet investissement permet une interprétation éclairée et concrète des phénomènes décrits et une vision critique sur la qualité des données analysées.

… et en mobilisant des études nationales, pour une mise en perspective

La correction manuelle et systématique des données erronées n'est évidemment pas possible ; un choix doit alors s'opérer pour trouver le meilleur compromis entre une qualité d'information acceptable, un nettoyage le moins chronophage possible et une analyse pertinente pour la mission. On peut dans ce cas faire appel à des études nationales, souvent produites par les SSM, pour mettre en perspective dans un contexte plus global les résultats obtenus localement.

Par exemple, dans le cadre des missions portant sur l'ASE, l'Igas s'appuie sur les analyses et l'expertise de la Drees. C'est grâce à ces statistiques qu'elle a pu détecter l'incomplétude des données disponibles dans le SI d'un département. En effet, l'augmentation du nombre d'enfants suivis par l'ASE retracé par le SI était inférieure à celle publiée par la Drees pour ce département, remettant ainsi en cause de manière plus générale l'exhaustivité des données dans les SI. Cette découverte a permis d'alerter l'ASE sur la non-exhaustivité des données dans ses systèmes, la conduisant à s'interroger sur ses procédures de saisie. L'Igas a quant à elle rectifié son approche pour prendre en compte cet élément dans ses analyses.

Analyser des pratiques locales pour les étendre à l'échelle nationale

> ***Au-delà des résultats d'évaluation à proprement parler, l'un des rôles de l'Igas consiste à réfléchir sur la possibilité pour les acteurs du terrain de mener leur propre analyse.***

Malgré les difficultés évoquées pour les exploiter, les données des SI locaux restent une source incontournable, et ce d'autant plus lorsqu'aucune information n'est disponible au niveau national. Les résultats sur une seule région ou un seul département ne sont certes pas aussi robustes qu'une évaluation faite sur l'ensemble du territoire, mais ils permettent d'émettre une hypothèse de tendance, de montrer la faisabilité d'une analyse, voire d'appuyer une recommandation. En effet, au-delà des résultats d'évaluation à proprement parler, l'un des rôles de l'Igas consiste à réfléchir sur la possibilité pour les acteurs du terrain de mener leur propre analyse. Plusieurs méthodes peuvent être testées et discutées, l'objectif étant de souligner la faisabilité de l'étude et l'avantage qu'auraient à tirer les acteurs du système à la réaliser sur des données plus complètes ou plus robustes.

La contribution des *data scientists* sur la mission « Les parcours des usagers de la sécurité sociale » (Fournier et al., 2025) illustre bien cette approche. Il s'agissait notamment d'étudier, à la suite de la mise en place d'une offre de service par l'Urssaf[9], si les contacts que les créateurs d'entreprise avaient avec un agent de l'organisme se traduisaient par un changement de comportement : amélioration de la qualité de leur déclaration et du paiement de leurs cotisations et/ou meilleur exercice de leur droit à une aide sociale. La Caisse nationale des Urssaf a transmis à l'Igas l'ensemble des informations sur les données financières et les contacts entrants (émanant des créateurs d'entreprise), mais elle n'avait pas la trace des contacts sortants (émanant des agents de l'Urssaf). Ces données permettaient donc d'avoir une vision exhaustive sur l'impact des contacts entrants, mais il subsistait une certaine frustration à ne pas pouvoir mesurer l'impact des contacts sortants. Cependant, le SI local de l'Urssaf en région Languedoc-Roussillon disposait de cette information, ce qui a permis de réaliser l'étude dans cette région. Les résultats sont certes limités à ce seul territoire, mais ils montrent la faisabilité de l'analyse et offrent une première estimation pour quantifier l'impact de ces appels. Ils démontrent aussi l'intérêt de tracer cette information, et donc d'étendre cette pratique à l'ensemble des Urssaf. C'est en effet un véritable enjeu pour les caisses régionales, qui pourraient ainsi mieux cibler les destinataires des appels émis par leurs conseillers.

Savoir aller chercher la donnée quand elle n'est pas recensée ou disponible…

Les SI locaux sont une véritable mine d'informations que l'Igas ne néglige pas, malgré toutes les difficultés inhérentes à leurs exploitations. Toutefois, certaines missions peuvent nécessiter de s'appuyer sur des données qui ne sont consolidées nulle part, en raison de la nouveauté ou du manque d'exploitation du sujet. L'Igas peut alors créer ses propres bases pour répondre au besoin spécifique des missions.

> ***L'Igas peut créer ses propres bases pour répondre au besoin spécifique des missions.***

Cette approche, bien que pragmatique, soulève des limites méthodologiques, notamment sur le plan de la représentativité des données et de la robustesse des analyses. Jusqu'à présent, l'Igas a fréquemment eu recours à des enquêtes de terrain pour pallier ces lacunes, une solution toutefois contraignante. Les résultats de telles enquêtes nécessitent un contrôle rigoureux pour garantir leur représentativité et leur redressement peut être chronophage. De plus, cette méthode sollicite les interlocuteurs locaux, déjà confrontés à une charge de travail importante, ce qui peut altérer la qualité et la disponibilité des données recueillies. L'Igas s'efforce désormais d'éviter autant que possible les enquêtes en utilisant des techniques comme le *web scraping*[10] (Lotfi et al., 2021) pour accéder à des données existantes. Cela réduit la sollicitation des acteurs locaux tout en optimisant les analyses avec des données exploitables et dont la fiabilité est acceptable.

9 Urssaf : union de recouvrement des cotisations de sécurité sociale et d'allocations familiales. L'Agence centrale des organismes de sécurité sociale (Acoss) et le réseau des Urssaf collectent et gèrent les ressources de la majorité des organismes de protection sociale.

10 Le *web scraping*, ou moissonnage en français, est une technique d'extraction automatisée de données de site web.

À titre d'exemple, dans le cadre de la mission « Lieux de vie et accompagnement des personnes âgées en perte d'autonomie », il convenait absolument de disposer d'une vision globale sur l'offre d'hébergement que proposent les résidences services. Ces dernières sont des structures privées non médicalisées, qui n'obéissent pas au code de l'action sociale et des familles. Elles n'appartiennent donc pas au champ sanitaire et social et ne sont pas recensées dans le répertoire FINESS[11] (Bensoussan et al., 2023). Toutefois, il s'agit d'un acteur important dans l'accueil des personnes âgées, qui tiendra probablement une place de plus en plus grande dans les prochaines années. Pour l'analyse et les projections réalisées, ces résidences devaient donc être prises en compte dans la capacité d'accueil au même titre que les établissements d'hébergement pour personnes âgées dépendantes (Ehpad) ou les résidences autonomie. Les acteurs dans le domaine étant très limités, deux sites web recensaient l'essentiel de l'offre disponible sur le territoire. Le *web scraping* a permis de constituer une base des résidences services ouvertes (ou avec une date d'ouverture prévisionnelle), avec leur lieu d'implantation, les logements disponibles et leur taille, le montant du loyer et les services proposés. Cette base a complété la connaissance sur l'offre d'accueil dans les départements et a été largement utilisée par la mission pour les projections d'accueil des personnes âgées par les différents acteurs du système. L'Igas a bien conscience de la fragilité de ces données, notamment sur les services offerts, dans la mesure où elles s'appuient sur des communications commerciales ; les constats et recommandations en tiennent évidemment compte. Néanmoins, connaître par département cette capacité d'accueil a enrichi substantiellement l'analyse.

… ou apprendre à s'en passer

> ***Le* web scraping *constitue un outil intéressant, mais il n'est malheureusement pas la solution idéale au problème de données non consolidées.***

Le *web scraping* constitue un outil intéressant, mais il n'est malheureusement pas la solution idéale au problème de données non consolidées. Il suppose de pouvoir répliquer la même requête sur un même site, ou bien de requêter un ensemble de sites de façon similaire. Si les données sont disponibles sur des sites différents, par exemple un site par département dans le cas des résidences services, il est impératif que tous les sites soient construits sur le même modèle. Si la requête se fait sur le même site, ce sont cette fois les outils de sécurité qui sont potentiellement un obstacle. En effet, une même requête répétée plusieurs fois peut être interprétée comme une cyberattaque et donc être rejetée par le serveur. Ces difficultés conduisent parfois l'Igas à s'adapter pour éclairer tout de même le décideur sur le sujet.

Ainsi, pour la mission « Évaluation de l'encadrement, de l'organisation et de la qualité des vacances adaptées organisées (VAO) », les inspecteurs souhaitaient dans un premier temps dresser un état des lieux des VAO. Une approche sur les sites régionaux ou départementaux aurait pu répondre au besoin, mais tous n'avaient pas l'information. La multiplicité de l'offre ne permettait pas, par ailleurs, une approche par organisme. Des techniques de *web scraping* ont alors été appliquées sur seulement deux organismes

11 FINESS est le répertoire des établissements sanitaires et sociaux.

bien implantés dans le secteur. L'objectif était d'obtenir l'ensemble des informations sur les séjours proposés aux personnes handicapées (prix, nombre de participants et d'accompagnants), pour les comparer ensuite à leur offre générique. Cette information a permis à la mission d'éclairer concrètement mais partiellement la situation des VAO (Leconte et Itier, 2024).

Le pôle *data* de l'Igas peut être amené à utiliser d'autres techniques d'extraction. Ainsi, le fait que beaucoup d'acteurs du champ social bénéficient d'un financement public, y compris des acteurs privés, se traduit par une place importante du contrôle dans les missions de l'Igas. La coopération des acteurs dans le contrôle ou simplement leur capacité à fournir les éléments utiles pour l'analyse sont alors un élément central des missions. Il peut arriver ainsi que les interlocuteurs transmettent des données numériques sous format PDF, par exemple des liasses fiscales. Or, les outils préformatés de transformation des fichiers ne donnent pas toujours des résultats satisfaisants, ou ils ne respectent pas la confidentialité s'ils doivent transiter par le web pour aller sur un outil d'intelligence artificielle. Dans ce cas, l'Igas a recours à un outil de reconnaissance optique de caractères (ROC ou OCR en anglais pour *optical character recognition*). Un tel outil permet d'extraire les informations importantes et de les stocker dans les fichiers au format ad hoc. Cette technique nécessite un investissement assez coûteux en temps et une structure uniforme des fichiers, mais elle permet de contourner la difficulté. À terme, les outils d'intelligence artificielle installés dans un environnement sécurisé devraient probablement répondre au besoin.

▶ Et maintenant ?

Après deux ans d'existence, marqués par l'exploitation de données dont la structure, la qualité, la fraîcheur et la disponibilité restent variables, l'intégration d'une approche *data* dans les missions de l'Igas s'intensifie significativement. Dans ce contexte, il devient impératif pour le pôle d'organiser la capitalisation des travaux déjà réalisés, afin d'en optimiser l'usage et d'en renforcer l'impact.

En premier lieu se pose la question de la maintenance et de l'actualisation des bases de données créées. Le besoin de mise à jour, mais aussi la pertinence et la complétude des données collectées, doivent être évalués. L'évolution du contexte sanitaire et social est un élément majeur qui influence directement les données et dont l'Igas va avoir besoin pour ses analyses. La mise à jour des bases construites par le pôle doit intégrer cette dimension.

Capitaliser au fur et à mesure des missions

> ***L'écueil consisterait à répliquer d'une mission à l'autre la même approche data, alors qu'il est primordial d'évoluer et de prendre du recul.***

L'Igas réalise régulièrement des missions de contrôle de certains organismes (organismes faisant appel à la générosité publique (OFAG), centres de formation d'apprentis (CFA), etc.) ou d'évaluation d'organismes lors d'un changement de directeur (centres hospitaliers universitaires (CHU), Dreets, etc.). Ces dernières missions sont appelées « T0 ». La récurrence des missions de

contrôle et d'évaluation permet de monter en compétence sur les données, d'améliorer la compréhension qu'en a l'inspection et l'usage qu'elle peut en tirer. Même si le thème de ces missions est identique, chacune doit être perçue comme un nouveau sujet : l'écueil consisterait en effet à répliquer d'une mission à l'autre la même approche *data*, alors qu'il est primordial d'évoluer et de prendre du recul.

Par exemple, dans le cadre du T0 d'un CHU, il est envisageable de systématiser l'extraction de certaines données financières ou d'activité, ainsi que le calcul par grand thème de certains indicateurs identifiés comme pertinents pour comprendre les forces et les voies d'amélioration de l'établissement : évolution des nombres de séances et de patients traités en dialyse ou en chimiothérapie, part de la chirurgie ambulatoire dans l'activité, etc. Cette approche est seulement un premier pas pour orienter des recherches plus poussées sur la situation globale de l'établissement, en prenant par exemple en compte d'autres facteurs comme la concurrence ou les changements de personnel. Il est impératif que les inspecteurs creusent chaque piste afin d'enrichir leur constat et de mettre en œuvre de nouvelles analyses. Ces dernières pourront à leur tour donner lieu à la définition d'un nouvel indicateur qui sera par la suite calculé systématiquement. C'est pourquoi les *data scientists* doivent échanger avec les inspecteurs en aval des missions, pour construire et enrichir avec eux les travaux qu'ils réalisent.

> ***C'est pourquoi les* data scientists *doivent échanger avec les inspecteurs en aval des missions, pour construire et enrichir avec eux les travaux qu'ils réalisent.***

Enfin, même lorsqu'elles ne sont pas récurrentes, les missions menées par l'Igas répondent aux besoins des décideurs pour orienter leurs politiques et sont en lien direct avec les problématiques de la société. De ce fait, sur une période donnée, plusieurs d'entre elles peuvent porter sur des problématiques communes. Il est donc important que le pôle sache mettre à profit ses travaux au-delà du champ de la mission elle-même, afin que les missions connexes puissent également en bénéficier. La prise en charge du grand âge qui a donné lieu à plusieurs missions sur les deux dernières années illustre bien cette problématique. Le fait que les analyses quantitatives aient été réalisées par la même équipe constitue un véritable atout. Connaître les bases disponibles et avoir déjà analysé une grande partie des données nécessaires à la mission permet de gagner en efficacité et surtout en qualité de l'analyse. Le *data scientist*, déjà averti sur la problématique, peut avoir une vision plus profonde du sujet et proposer à la mission des pistes d'analyse qu'il n'aurait pas identifiées sans ce travail préalable.

L'Igas doit oser des opérations innovantes

L'Igas réalise des missions sur un thème précis et ne doit pas nécessairement répondre à des besoins de représentativité nationale. Ainsi, obéissant à moins de contraintes que le système statistique public, le pôle *data* peut réaliser des projets pilotes en collaboration avec les SSM avant que ces derniers n'avancent sur le sujet.

Par exemple, les missions de l'Igas intègrent de plus en plus la notion de parcours. Il peut s'agir de parcours professionnels, comme par exemple dans la mission évoquée plus haut sur les temps partiels contraints, mais aussi de parcours regroupant des thématiques plus variées, nécessitant l'interrogation et la mise en cohérence de plusieurs sources émanant de producteurs différents. L'Igas n'est pas isolée dans cette volonté d'approcher les problématiques des citoyens dans leur globalité. Elle partage ce désir avec ses principaux partenaires de la statistique publique comme la Drees et la Dares, mais ces derniers opèrent dans un cadre plus contraint sur la méthode et doivent s'assurer d'une représentativité à l'échelle nationale.

> ***Le positionnement du pôle* data *au sein d'une inspection et la connaissance précise qu'il a des sujets portant sur le champ social doivent l'inciter à innover dans ses analyses.***

Ainsi, dans le cadre de ses missions, l'Igas s'est associée aux travaux menés par les services de la statistique publique pour étudier la faisabilité d'un croisement de fichiers. L'objectif était de rapprocher les données sur la formation et l'emploi issues de la base Minimas sociaux, droits d'assurance chômage et parcours salariés (Midas) avec celles de l'ASE sur le parcours des enfants passés par leurs services. Ce travail est encore à ses débuts : l'Igas propose d'expérimenter ce rapprochement sur quatre services départementaux d'ASE qu'elle a étudiés. Un travail entre les producteurs de données, le CASD et l'Igas est en cours. Il pourrait permettre d'étudier localement le devenir professionnel de ces enfants, dont on ne sait encore que très peu de choses. Mais aussi, il pourrait donner un aperçu des difficultés dans la mise en relation des différentes bases. L'avantage d'expérimenter ce rapprochement de sources dans le cadre de la mission est qu'il facilite l'accès direct aux acteurs du terrain, ce qui permet de lever des doutes sur les données locales et d'améliorer la compréhension des parcours. L'analyse de l'Igas ne répondrait pas à l'ensemble des standards et des exigences de la statistique publique, notamment en matière de représentativité. Ce serait une première ébauche, préalable à un travail d'analyse plus complet et représentatif que pourraient faire les SSM.

Il ne s'agit là que d'un cas particulier. Plus largement, le positionnement du pôle *data* au sein d'une inspection et la connaissance précise qu'il a des sujets portant sur le champ social doivent l'inciter à innover dans ses analyses.

C'est technique, pas magique !

La diversité des missions, la variété des sujets abordés et la richesse des données disponibles font du travail du *data scientist* à l'Igas une expérience à la fois unique et passionnante. Lorsqu'il évolue dans un cadre bien balisé, son travail ressemble à des analyses statistiques classiques. Mais la réalité est souvent plus nuancée : contraintes de temps, qualité des données parfois aléatoire, etc. Le *data scientist* doit alors naviguer avec pragmatisme, s'éloignant parfois des canons de l'orthodoxie statistique pour orienter ses analyses de manière pertinente. Le résultat final, lui, doit rester limpide et accessible à tous, en masquant la complexité et les contorsions intellectuelles qui ont jalonné le processus.

Encore méconnu pour certains, le rôle du *data scientist* pourrait passer, du point de vue des utilisateurs, pour une « moulinette » mystérieuse produisant des résultats sans révéler ses secrets. Pourtant, après deux ans d'existence, le pôle *data* de l'Igas a su démontrer son utilité grâce à une collaboration étroite et constante avec les inspecteurs. Cette coopération a non seulement révélé le potentiel de la *data science* pour les missions de l'Igas, mais aussi mis en lumière ses contraintes et ses limites, en particulier celles liées aux données. Parce qu'en *data science* comme ailleurs, il n'y a pas de magie : juste des données, des algorithmes, du sens critique et du travail.

▶ Bibliographie

- BENSOUSSAN, Johanna, BIZINGRE, Joël et COURVALIN, Nathalie, 2023. FINESS, le répertoire des établissements de santé. In : *Courrier des statistiques*. [en ligne]. 11 décembre 2023. Insee. N° N10, pp. 71-92. [Consulté le 4 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7722095?sommaire=7722116*.

- BOURLANGE, Danielle, BRUNET, François, CHIGNARD, Simon et EIDELMAN, Alexis, 2021. Évaluation des besoins de l'État en compétences et expertises en matière de donnée. In : *Rapport de la DINUM et de l'Insee*. [en ligne]. Juin 2021. [Consulté le 4 mars 2025]. Disponible à l'adresse : *https://www.vie-publique.fr/rapport/281669-besoins-de-etat-en-competences-expertises-en-matiere-de-donnees*.

- BRUNET, François, ROTH, Nicole et SCHAPIRA, Irina, 2023. Utilisation des données de la Déclaration sociale nominative (DSN) à des fins de statistiques publiques ou de pilotage. In : *Rapport de l'inspection générale de l'Insee / Rapport de l'Igas*. [en ligne]. Avril 2023. [Consulté le 18 mars 2025]. Disponible à l'adresse : *https://www.igas.gouv.fr/utilisation-des-donnees-de-la-declaration-sociale-nominative-dsn-des-fins-de-statistiques-publiques*.

- COTTON, Franck et HAAG, Olivier, 2023. L'intégration des données administratives dans un processus statistique – Industrialiser une phase essentielle. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp. 104-125. [Consulté le 4 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635829?sommaire=7635842*.

- COUDIN, Élise et ROBERT, Aude, 2024. Les statistiques sur les causes de décès – Classer et coder… dans la classification internationale des maladies. In : *Courrier des statistiques*. [en ligne]. 16 décembre 2024. Insee. N° N12, pp. 27-50. [Consulté le 4 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8264550?sommaire=8264562*.

- DIALLO, Cheikh Tidiane, MOREL-JEAN, Constance et SARRON, Clotilde, 2024. L'aide sociale départementale : bénéficiaires, dépenses, financement, personnel – Édition 2024. In : *Les dossiers de la DREES*. [en ligne]. 6 novembre 2024. DREES. N° 124. [Consulté le 4 mars 2025]. Disponible à l'adresse : *https://www.drees.solidarites-sante.gouv.fr/publications-communique-de-presse/les-dossiers-de-la-drees/241106_DD_aide-sociale-departementale*.

- EMMANUELLI, Julien, FROSSARD, Jean-Baptiste et VINCENT, Bruno, 2024. Lieux de vie et accompagnement des personnes âgées en perte d'autonomie : les défis de la politique domiciliaire, se sentir chez soi où que l'on soit. In : *Rapport de l'Igas*. [en ligne]. 29 mars 2024. [Consulté le 4 mars 2025]. Disponible à l'adresse : *https://igas.gouv.fr/Lieux-de-vie-et-accompagnement-des-personnes-agees-en-perte-d-autonomie-les*.

- FOURNIER, Valentine, GROSSE, Alexandre, VILBOEUF, Laurent et VINCENT, Bruno, 2025. Les parcours des usagers de la Sécurité sociale : Comment mieux accompagner les moments importants de la vie ? In : *Rapport de l'Igas*. [en ligne]. 8 avril 2025. [Consulté le 14 avril 2025]. Disponible à l'adresse : *https://www.igas.gouv.fr/les-parcours-des-usagers-de-la-securite-sociale-comment-mieux-accompagner-les-moments-importants-de-la-vie*.

- GADOUCHE, Kamel, 2019. Le Centre d'accès sécurisé aux données (CASD), un service pour la data science et la recherche scientifique. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 76-92. [Consulté le 4 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254227?sommaire=4254170*.

- HUMBERT-BOTTIN, Élisabeth, 2018. La déclaration sociale nominative – Nouvelle référence pour les échanges de données sociales des entreprises vers les administrations. In : *Courrier des statistiques*. [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 25-34. [Consulté le 4 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/3647025?sommaire=3647035*

- LECONTE, Thierry et ITIER, Christophe, 2024. Vacances organisées pour adultes handicapés : état des lieux et leviers d'amélioration. In : *Rapport de l'Igas*. [en ligne]. 2 juillet 2024. [Consulté le 4 mars 2025]. Disponible à l'adresse : *https://igas.gouv.fr/vacances-organisees-pour-adultes-handicapes-etat-des-lieux-et-leviers-damelioration.*

- LOTFI, Chaimaa, SRINIVASAN, Swetha, ERTZ, Myriam et LATROUS, Imen, 2021. Web Scraping Techniques and Applications: A Literature Review. In : *SCRS Conference Proceedings on Intelligent Systems*. [en ligne]. Janvier 2021. Pp. 381-394. [Consulté le 4 mars 2025]. Disponible à l'adresse : *https://www.researchgate.net/publication/367719780_Web_Scraping_Techniques_and_Applications_A_Literature_Review*.

- MAGNIER, Antoine et VIOSSAT, Louis-Charles, 2024. Temps partiel et temps partiel contraint : des inflexions possibles pour un cadre rénové. In : *Rapport de l'Igas*. [en ligne]. 17 décembre 2024. [Consulté le 4 mars 2025]. Disponible à l'adresse : *https://www.igas.gouv.fr/temps-partiel-et-temps-partiel-contraint-des-inflexions-possibles-pour-un-cadre-renove*.

Le code officiel géographique

La géographie sans les cartes

Joachim Clé, Frédéric Minodier**, Violaine Simon*** et Pierre Vernédal*****

Officialisé à l'âge de 60 ans (il en a maintenant plus de 80), le code officiel géographique décrit la France et plus que la France : c'est tout à la fois un répertoire et une nomenclature de territoires. Construit à l'origine pour les travaux statistiques et pour certaines tâches administratives, il fait aujourd'hui partie des neuf jeux de données de référence du « service public de la donnée », créé par la loi du 7 octobre 2016 pour une République numérique. Pourtant, les informations qu'il contient sont relativement élémentaires (un code et un libellé) et sa volumétrie est loin d'être massive, le plus gros contingent correspondant aux quelque 35 000 communes de la République. Il n'en reste pas moins que ce code, assis sur des textes réglementaires et servi par un travail minutieux de veille et de coopération entre administrations, est particulièrement utile, ses usages allant de la tenue de répertoires d'individus à l'industrie ou au monde de la santé. Cette production ancienne reste moderne, notamment au travers de sa diffusion par interface de programmation applicative et de ses évolutions régulières, pour décrire toujours mieux le territoire.

Made official at the age of 60 (it is now over 80), the official geographical code describes France and more than France: it is both a register and a classification of territories. Originally built for statistical work and some administrative tasks, it is now one of the nine reference datasets of the 'public data service', created by the Law of 7 October 2016 for a Digital Republic. Yet its content is relatively basic (a code and a label) and its volume is far from being massive, with the largest number of records corresponding to the Republic's 35,000 or so municipalities. However, this code, based on regulatory texts and made possible by meticulous monitoring and cooperation between administrations, is particularly useful, with uses ranging from keeping registers of individuals to industry and the health sector. This old production remains modern, particularly through its dissemination by application programming interface and its regular developments, to describe the territory ever better.

* Expert en information géographique, Direction régionale du Centre-Val de Loire, Insee.
joachim.cle@insee.fr

** À la date de la rédaction, chef de la division Méthodes et référentiels géographiques, DMCSI, Insee.
frederic.minodier@insee.fr

*** Chargée de travaux sur les référentiels géographiques, Direction régionale du Centre-Val de Loire, Insee.
violaine.simon@insee.fr

**** *Chef du pôle Référentiels géographiques, Direction régionale du Centre-Val de Loire, Insee.*
pierre.vernedal@insee.fr

Nichés au cœur du numéro d'inscription au répertoire des personnes physiques ou numéro de sécurité sociale (Espinasse et Roux, 2022), cinq caractères désignent votre lieu de naissance. Et non, ce n'est pas le code postal[1] qui est utilisé pour cela, c'est le code officiel géographique. Celui-ci avait déjà plus de vingt ans d'usage quand le code postal est apparu dans les années 1960.

▶ Un répertoire de territoires

Le code officiel géographique (COG – prononcé « coj ») est né le 1er octobre 1943 et a traversé depuis lors l'histoire de la statistique publique[2]. Selon le préambule de la deuxième édition du COG publiée en 1954 (Insee, 1954), il a été « établi pour permettre l'exécution des travaux statistiques et de certaines tâches administratives (telles que la détermination du numéro national d'identité des personnes physiques) ». C'est un ensemble de listes dont il décrit les évolutions : celles des communes, cantons, arrondissements départementaux, départements, régions, collectivités territoriales ayant les compétences départementales, collectivités et territoires français d'outre-mer, pays et territoires étrangers[3]. C'est donc en ce sens un répertoire. Un répertoire qui vit au rythme des modifications administratives (naissances, disparitions, fusions ou scissions de territoires, etc.).

> ***Même si sa population d'intérêt est de taille bien plus limitée, le COG s'apparente donc aux répertoires régaliens que sont Sirene pour les entreprises et les établissements et le RNIPP pour les personnes physiques.***

D'un répertoire, il possède les cinq propriétés (Bizingre et al., 2013 ; Rivière, 2022). D'abord celle d'**unité de sens** : le COG a un objet clair, les différents échelons administratifs du territoire. Puis celle de **stabilité** : les échelons administratifs évoluent peu, si bien que l'on dispose d'un historique complet de ces derniers depuis 1943. La propriété de **centralité** est moindre que pour les répertoires de personnes physiques ou d'entreprises, dans la mesure où le COG n'intervient pas dans le processus administratif de modification des territoires. Il en est cependant une chambre d'enregistrement : chaque mouvement dans le COG est relié à un acte administratif, tout comme pour le répertoire des établissements sanitaires et sociaux FINESS (Bensoussan et al., 2023). Quatrième propriété, la **qualité** du COG est entretenue et reconnue, ce qui lui procure une place de choix dans la production statistique, notamment pour le recensement de la population. Enfin, dernière propriété, l'**interopérabilité**, c'est-à-dire l'ouverture et l'accessibilité : elle est aujourd'hui assurée par une large mise à disposition du code au moyen de fichiers[4] et d'une interface de programmation applicative dédiée (ou API pour *application programming interface* en anglais)[5] (Mauguin et Sagnes, 2024).

1 Le code postal est attribué par La Poste. Il identifie un bureau distributeur de courrier.
2 René Carmille, directeur du Service national des statistiques au moment de la création du COG, en expose clairement l'idée dans son ouvrage sur la mécanographie dans les administrations (Carmille, 1942).
3 Géographie administrative, *https://www.insee.fr/fr/information/8064273*.
4 Les fichiers constitutifs du COG : *https://www.insee.fr/fr/information/2560452*.
5 Une API est une interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service, afin d'échanger des données et des fonctionnalités.

Même si sa population d'intérêt est de taille bien plus limitée, le COG s'apparente donc aux répertoires régaliens que sont Sirene[6] pour les entreprises et les établissements et le RNIPP[7] pour les personnes physiques. Comme ses deux alter ego, il attribue de façon univoque un identifiant pérenne à tout nouvel enregistrement, et par là constitue la référence pour le lien entre les « traits d'identité » qui le désignent habituellement (par exemple, pour une commune, son nom) et l'identifiant (son code officiel géographique). C'est un répertoire socle dont l'utilisation dépasse le simple usage statistique. Ainsi, le numéro d'inscription au répertoire (NIR), communément appelé numéro de sécurité sociale, comprend le code géographique du lieu de naissance de la personne. Grâce à ce code, il est donc possible de retrouver le libellé du lieu de naissance de la personne en vigueur à la date de sa naissance, quelles que soient les évolutions qui ont pu affecter le territoire. Pour l'anecdote, avant la création du répertoire Sirene en 1973, l'identifiant administratif des établissements incluait, quant à lui, les codes du département et de la commune d'implantation.

▶ Une nomenclature du territoire français…

> ***Il aura ainsi fallu soixante ans pour que le code officiel devienne « officiellement » officiel.***

Le COG est aussi une nomenclature : l'arrêté du ministre de l'Économie, des Finances et de l'Industrie du 28 novembre 2003 relatif au code officiel géographique[8] l'institue en tant que « nomenclature des collectivités territoriales et des circonscriptions administratives de la France ». Il aura ainsi fallu soixante ans pour que le code officiel devienne « officiellement » officiel (Lang, 2003). En tant que nomenclature, il articule entre eux les différents éléments qui le constituent (régions, départements, communes, etc.).

De même que la nomenclature d'activités française (NAF)[9] comporte cinq niveaux, de la section à la sous-classe, le COG distingue divers types d'objets dont l'emboîtement permet d'avoir différents niveaux d'analyse tout en permettant le passage d'un niveau à l'autre : les régions contiennent les départements, qui contiennent les arrondissements et les cantons, dans lesquels on trouve les communes, briques élémentaires du COG (***figure 1***).

À noter que la nature des objets n'est pas décidée par l'Insee mais par la loi : par exemple, en 1982, les régions deviennent des collectivités territoriales et, à ce titre, un nouveau type d'objet fait son entrée dans le COG (Lang, 2003). Plus récemment, la loi a créé la collectivité territoriale à compétences départementales avec l'institution au 1er janvier 2015 de la métropole de Lyon[10].

6 Système informatique pour le répertoire des entreprises et des établissements, *https://www.insee.fr/fr/metadonnees/source/serie/s1020*.
7 Répertoire national d'identification des personnes physiques, *https://www.insee.fr/fr/information/5019311*.
8 Voir les références juridiques en fin d'article.
9 Nomenclature d'activités française, *https://www.insee.fr/fr/information/2406147*.
10 Pour la codification des collectivités territoriales à compétences départementales et à statut particulier, voir : *https://www.insee.fr/fr/information/7929497*.

▶ Figure 1 - Des partitions de la France

Chiffres clés au 1er janvier 2025	Exemple d'Aubusson* et des territoires qui englobent cette commune		
	Niveau géographique	*Code*	*Libellé*
18 régions	Région	75	Nouvelle-Aquitaine
101 départements	Département	23	Creuse
333 arrondissements	Arrondissement	231	Aubusson
2 054 cantons	Canton	2302	Aubusson
34 875 communes	Commune	23008	Aubusson

Champ : France métropolitaine et départements et régions d'outre-mer.

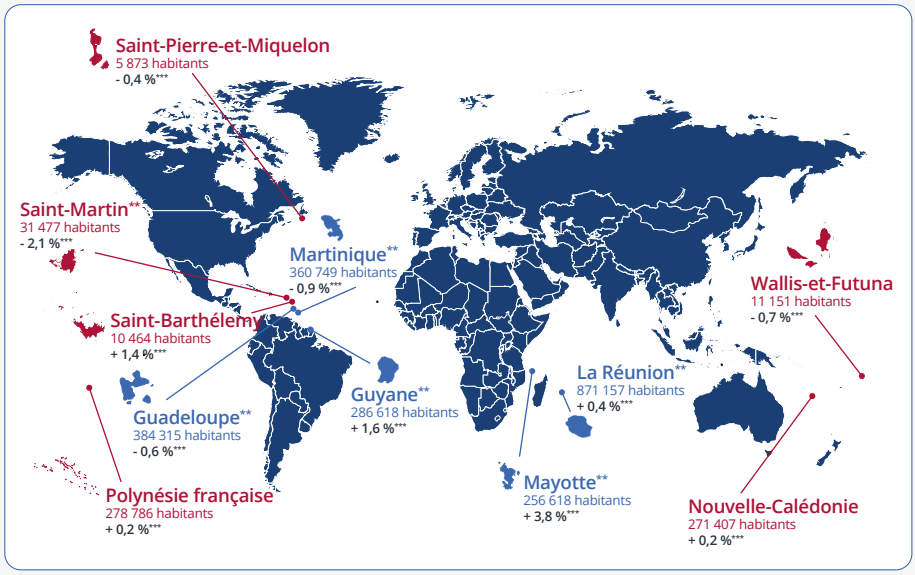
* Plus d'informations sur la page : *https://www.insee.fr/fr/metadonnees/geographie/commune/23008-aubusson*.

Comme toute nomenclature, le COG a vocation à représenter l'ensemble de ce qu'il décrit et offre ainsi une partition du territoire souverain de la République française[11] au travers des collectivités territoriales et des circonscriptions administratives. Les collectivités territoriales sont définies dans le titre XII éponyme de la Constitution[12], à savoir « les communes, les départements, les régions, les collectivités à statut particulier et les collectivités d'outre-mer régies par l'article 74 ». L'article 72-3 indique que, pour les populations d'outre-mer, la Guadeloupe, la Guyane, la Martinique, La Réunion, Mayotte, Saint-Barthélemy, Saint-Martin, Saint-Pierre-et-Miquelon, les îles Wallis et Futuna et la Polynésie française sont régis par l'article 73 pour les DROM[13], soit les cinq premiers, et par l'article 74 pour les autres collectivités ; le statut de la Nouvelle-Calédonie est régi par le titre XIII et celui des Terres australes et antarctiques françaises (TAAF) et de La Passion-Clipperton par la loi (***figure 2***). Les cantons et les arrondissements (départementaux) font historiquement partie intégrante du COG (***encadré 1***).

Le COG décrit donc l'ensemble du territoire français, tout en se gardant bien de définir ce qu'on entend par France. Cette question de terminologie est réapparue périodiquement dans les publications statistiques, notamment au fur et à mesure de l'intégration des DOM[14] dans l'ingénierie statistique (Insee et al., 2023) : le total France métropolitaine a cédé la place au total France entière (ou France hors Mayotte) pour marquer l'inclusion des DOM, évoluant aujourd'hui simplement vers France. On le voit, le terme France en statistique[15] ne désigne pas l'ensemble de la République française mais le regroupement des départements régis par les articles 72 et 73 de la Constitution. Légalement, la seule responsabilité du service statistique public à s'étendre sur l'ensemble du territoire

11 À l'exception bien sûr des ambassades étrangères en France et françaises à l'étranger.
12 Voir les références juridiques en fin d'article.
13 DROM : département et région d'outre-mer.
14 DOM : département d'outre-mer.
15 L'article 9 de la loi de 1951 précise qu'elle est « applicable dans les territoires d'outre-mer et les territoires associés ».

▶ Figure 2 - Les Outre-mer*



Légende

Départements et régions d'outre-mer

Collectivités d'outre-mer

* Les Terres australes et antarctiques françaises, l'île de La Passion-Clipperton et l'Île des Faisans ne figurent pas sur la carte. Ces territoires ultramarins ne comptent aucun habitant permanent.
** Région ultrapériphérique (RUP). La métropole et les six RUP constituent le champ géographique des engagements européens en matière statistique.
*** Taux d'évolution annuel de la population (en moyenne sur cinq ans).
Source : Insee, recensements de la population 2015 et 2021 (Guadeloupe, Martinique, Guyane, La Réunion, Saint-Pierre-et-Miquelon, Saint-Barthélemy, Saint-Martin), 2012 et 2017 (Mayotte), 2014 et 2019 (Nouvelle-Calédonie), 2018 et 2023 (Wallis-et-Futuna), 2017 et 2022 (Polynésie française) ; extrait de (Insee et al., 2023).

est le recensement de la population (et le calcul des populations de référence). Les engagements européens sont (un tout petit peu) plus larges puisque la partie française de l'île de Saint-Martin a le statut européen de région ultrapériphérique.

▶ ... et un peu plus

L'article 1 de l'arrêté du 28 novembre 2003 ne s'arrête pas au territoire français. Il institue le COG comme « la nomenclature [...] des pays et territoires étrangers ». Ainsi, le COG offre une description du monde avec l'ensemble des pays et des territoires étrangers ayant leur propre code géographique[16]. Cette section du COG existe depuis la première édition et a connu une existence tourmentée (Lang, 2003). Récemment rénovée, elle retrace aujourd'hui l'évolution du monde pendant les quatre-vingts dernières années (***encadré 2***).

16 Il convient de noter que la mention d'un pays ou territoire étranger dans le COG ne vaut pas sa reconnaissance diplomatique par la République française (voir *https://www.diplomatie.gouv.fr/fr/*).

▶ Encadré 1. Les objets du COG, les événements enregistrés et les textes réglementaires associés

Type d'objet géographique	Nature de l'événement	Texte actant l'événement
Commune	Création ou extension de commune nouvelle	Arrêté préfectoral
	Rétablissement de commune	Arrêté préfectoral
	Changement de nom	Décret
	Modification des limites territoriales à la suite d'un échange de parcelles, sans incidence sur les découpages administratifs supracommunaux	Arrêté préfectoral
	Modification des limites territoriales à la suite d'un échange de parcelles affectant les limites territoriales des cantons, départements ou régions	Décret
Arrondissement	Création, suppression, changement de nom, transfert de chef-lieu	Décret
	Modification des limites territoriales	Arrêté préfectoral
Collectivité territoriale à compétences départementales / Collectivité territoriale à statut particulier	Création, changement de nom, transfert de chef-lieu, modification de statut juridique, modification des limites territoriales	Loi, décret
Département	Changement de nom, transfert de chef-lieu, modification des limites territoriales	Décret
Région	Changement de nom, transfert de chef-lieu, modification des limites territoriales	Loi, décret
Collectivité et territoire d'outre-mer	Création, suppression, changement de nom, transfert de chef-lieu, modification des limites territoriales	Loi, décret
District administratif (Terres australes et antarctiques françaises)	Création, suppression, changement de nom, modification des limites territoriales	Loi, décret
Circonscription territoriale (îles Wallis et Futuna)	Création, suppression, changement de nom, transfert de chef-lieu, modification des limites territoriales	Loi, décret
Découpages territoriaux diffusés en plus de ceux formant le périmètre réglementaire du COG		
Canton	Création, suppression, changement de nom, modification des limites territoriales	Décret
Commune associée*	Suppression	Arrêté préfectoral
	Passage du régime de fusion association au régime de fusion simple	Arrêté préfectoral
	Transformation de commune associée en commune déléguée	Délibération de conseil municipal
	Rétablissement de commune associée supprimée sous forme de commune déléguée**	Arrêté préfectoral
Commune déléguée	Création (dans le cadre de la création d'une commune nouvelle)	Arrêté préfectoral
	Suppression	Délibération de conseil municipal
	Maintien de commune déléguée préexistante à la création d'une commune nouvelle issue de deux fusions successives	Délibération de conseil municipal
Arrondissement municipal (Paris, Lyon, Marseille)	Création, suppression, modification des limites territoriales	Loi

* *Au sens de la loi n° 71-588 du 16 juillet 1971 sur les fusions et regroupements de communes, dite « Loi Marcellin ». Voir les références juridiques en fin d'article.*

** *Dispositif temporaire créé par la loi n° 2019-809 du 1er août 2019 pour une année après la publication de la loi. Voir les références juridiques en fin d'article.*

> ***Le COG offre une description du monde avec l'ensemble des pays et des territoires étrangers.***

La définition du contenu du COG reste néanmoins assez évasive et la liste des zonages qu'il doit comporter n'est pas complètement explicite. Historiquement, toutes les éditions ont comporté les découpages suivants : les communes (et les arrondissements municipaux de Paris, Lyon, Marseille), les cantons, les arrondissements départementaux, les départements, les régions (à compter de leur création en 1982), les territoires d'outre-mer français (avec des statuts variables dans le temps) et les pays et territoires étrangers (Lang, 2017).

Ces zonages ont comme point commun l'attribution d'un code par l'Insee. Les communes, départements et régions sont des collectivités territoriales définies comme telles dans le Code général des collectivités territoriales (CGCT)[17]. Les cantons y figurent toujours, même si ce découpage ne revêt plus aujourd'hui qu'une valeur électorale depuis la réforme de la carte cantonale (loi n° 2013-403 du 17 mai 2013[18]).

En ce qui concerne l'outre-mer, les cinq départements et les collectivités se partagent les codes 97 et 98 et, enfin, les pays et territoires étrangers commencent par 99. Les codes en 97, initialisés par les quatre premiers DOM en 1946, ont évolué de manière discontinue au fil du temps, rompant la logique initiale : Saint-Pierre-et-Miquelon (975), un temps classé en DOM, est redevenu une collectivité d'outre-mer en 1985 ; Mayotte reçoit le code 976 en 2007 en tant que collectivité d'outre-mer avant d'intégrer la liste des départements au 1er janvier 2012 ; les îles de Saint-Martin et de Saint-Barthélemy sont séparées de la Guadeloupe en 2007.

▶ Encadré 2. Une nomenclature rénovée des pays et territoires étrangers

La publication du millésime 2024 a été l'occasion de retracer un historique complet des événements enregistrés : au lieu d'une succession de photos, il est désormais possible de faire le lien entre deux périodes pour reconstituer l'histoire d'un pays ou d'un territoire. À titre d'exemple, la scission de l'Allemagne en 1949 a entraîné la création d'un code pour l'Allemagne de l'Ouest et d'un code pour l'Allemagne de l'Est, codes qui ont pris fin en 1990 avec la réunification du pays.

Jusqu'en 2023, l'Insee proposait uniquement un fichier millésimé des codes des pays et territoires étrangers. Celui-ci présentait de nombreux défauts : territoires obsolètes, libellés erronés, plusieurs libellés pour un même code. De plus, les utilisateurs pouvaient difficilement effectuer une recherche historique sur ces codes et obtenir le bon libellé de pays ou territoire selon l'année. Il était donc compliqué d'associer le code pays présent dans le NIR d'une personne née à l'étranger à son réel pays de naissance.

Les codes des pays et territoires étrangers sont présents depuis la première édition du COG en 1943. De fait, il apparaissait nécessaire que ces codes intègrent la base de données géographiques de l'Insee, à l'instar des zonages français.

Les différentes éditions du COG ont été analysées pour lister les codes commençant par 99 et reconstituer les liens entre eux, que ce soient des liens temporels (division ou réunification de territoires par exemple) ou des liens d'appartenance (territoires dépendants d'un pays). On dispose in fine d'une liste de pays et de territoires, avec leurs dates de début et de fin d'existence dans le COG, leur statut, leurs libellés courts et longs, ainsi que d'une chronologie d'événements retraçant la géopolitique de ces quatre-vingts dernières années. Le concours de la Commission nationale de toponymie et du ministère de l'Europe et des Affaires étrangères a été sollicité pour l'expertise toponymique des libellés et des dates retenues.

L'Insee a ainsi pu publier en 2024 un fichier millésimé rénové des pays et territoires étrangers, ainsi que les liens codes-pays de 1943 à nos jours.

17 Voir les références juridiques en fin d'article.
18 Voir les références juridiques en fin d'article.

▶ Une commune, une région, c'est d'abord un code associé à un libellé…

> ***La représentation géométrique des territoires est du ressort de l'Institut national de l'information géographique et forestière (IGN).***

Mais, pour un territoire donné, quelles informations contient précisément le COG ? Un minimum de choses, puisque chaque objet n'a que deux caractéristiques : son code et son libellé. En particulier, la géométrie, c'est-à-dire le tracé des contours du territoire, n'est pas intégrée. C'est donc la géographie sans la géométrie et par conséquent sans cartes ! La représentation géométrique des territoires[19] est en effet du ressort de l'Institut national de l'information géographique et forestière (IGN). Celui-ci produit en particulier un fond cartographique des limites administratives françaises (Admin Express COG) calé une fois par an sur le COG millésimé (IGN, 2024).

La correspondance entre code et libellé est intéressante, car un code permet de fournir le libellé de la commune idoine. L'inverse est intéressant également : retrouver le code d'une commune à partir de son nom. Cependant, l'homonymie est assez fréquente : 10 % des communes portent le même nom qu'une autre. Heureusement, elles se situent le plus souvent dans des départements différents. Il existe néanmoins plusieurs cas de parfaite homonymie dans un département : outre les deux communes de Bors (16052 et 16053) dans le département de la Charente et celles de Castillon (64181 et 64182) dans les Pyrénées-Atlantiques, qui peuvent être distinguées par la mention de leur canton, les deux communes de Château-Chinon (58062 et 58063) dans la Nièvre, appartenant au même canton, sont repérées par les suffixes « ville » et « campagne » !

Le choix des dénominations n'est pas fait ex nihilo par l'Insee, contrairement à la codification. Pour les communes, le changement de nom trouve nécessairement son origine dans un décret, selon une procédure bien établie (article L2111-1 du CGCT) (***encadré 3***), et le choix du nom suit les recommandations émises par la Commission nationale de toponymie (CNT, 2021).

▶ … et toute une histoire !

Écart par rapport à un répertoire stricto sensu, code et libellé ne sont pas stables dans le temps pour un même objet géographique : dans le cas d'une fusion de communes, on peut garder le même code et le même libellé alors que l'objet change de manière significative. Cela est gênant pour un répertoire de ne pas avoir d'identifiant propre, non ? Qu'à cela ne tienne, les événements administratifs affectant les territoires étant rigoureusement suivis dès l'édition de 1971, il est possible de reconstituer un historique complet depuis 1943 et donc de savoir précisément ce que désigne un code pour une année donnée et de calculer les statistiques sur le bon périmètre (ouf !).

19 Représentation qui peut varier selon l'échelle, la qualité des informations pour créer le contour, etc.

▶ Encadré 3. Le COG, garant du nom des communes françaises

Changer le nom d'une commune n'est pas une simple formalité : selon une note d'information de la direction générale des collectivités locales (DGCL)*, le changement doit être « justifié par le souci de mettre le nom officiel de la commune en accord avec un usage différent mais suffisamment ancien et constant ou par celui de mettre fin à de véritables risques de confusion avec d'autres communes. Ne sont pas admises les modifications fondées sur des considérations de simple publicité touristique ou économique ». La Commission nationale de toponymie fournit un ensemble de recommandations pour décider du nom d'un lieu.

Mieux encore, cette note précise : « Est considéré comme officiel le nom de la commune tel qu'il apparaît dans le code officiel géographique ». Toute modification du nom de la commune, que ce soit « la substitution d'un nom à un autre, mais aussi les additions de noms ou les simples rectifications d'orthographe », est ainsi entérinée par décret (article L2111-1 du CGCT) avant d'être intégrée dans le COG.

La commune adresse sa demande à la préfecture de département. Le préfet, le conseil départemental et le directeur des archives départementales donnent ensuite leur avis sur cette requête. Cette dernière est enfin analysée au cours de la réunion d'examen des changements de nom de communes, organisée par la DGCL une ou deux fois par an. Cette réunion rassemble les représentants d'organismes concernés par les questions de toponymie française et l'évolution du nom des communes. L'Insee y participe en tant que gestionnaire du COG, ce dernier permettant en outre de reconstituer l'historique des modifications de dénominations communales depuis 1943. Près de 1 400 communes ont ainsi officiellement changé de toponyme, souvent en apportant des précisions au nom originel : elles étaient 14 en 2023, 9 en 2024, 8 en 2025.

La procédure est plus légère pour fixer le nom des communes nouvelles : selon l'instruction de 2017**, les conseils municipaux des communes fondatrices décident par délibérations concordantes du nom de la commune nouvelle. Le préfet de département s'assure du respect des règles de graphie toponymique et de l'absence d'homonymie. La dénomination choisie est ensuite rendue officielle par l'arrêté préfectoral portant création de la commune nouvelle.

* *Voir les références juridiques en fin d'article.*

** *Voir les références juridiques en fin d'article.*

> ***Dans le cas d'une fusion de communes, on peut garder le même code et le même libellé alors que l'objet change de manière significative.***

Ces événements (***encadré 1***) ont des natures et incidences assez diverses, allant de la création d'une commune au « simple » changement d'adresse de la mairie, lequel peut entraîner un changement de code de la commune. Mais tous font l'objet de textes dûment enregistrés en vertu de dispositions réglementaires, à l'instar des structures du répertoire FINESS. Le plus haut niveau est la loi, avec par exemple le changement de nom récent de l'île de Clipperton, aujourd'hui île de La Passion-Clipperton (loi 3DS du 21 février 2022[20]). Toutefois, la majeure partie des événements intéressant le COG sont à chercher auprès des préfectures de départements.

▶ Un outil pour le recensement de la population…

Le suivi de ces événements est particulièrement important pour l'organisation du recensement de la population et la publication de ses résultats, notamment les populations de référence authentifiées.

20 Voir les références juridiques en fin d'article.

À l'origine, les dénombrements de la population organisés par le ministère de l'Intérieur ont nécessité des listes de départements, arrondissements départementaux, cantons et communes, c'est-à-dire le cœur du COG. Pour les besoins de diffusion du recensement notamment, le référentiel des nomenclatures géographiques[21] de l'Insee accueille maintenant un spectre plus large que les seuls objets du COG. Ainsi, il inclut également les quartiers prioritaires de la politique de la ville, les communes associées[22] et les communes déléguées[23], dont les populations de référence font partie intégrante de la publication authentifiée par décret[24] chaque fin d'année.

> ***La publication d'une nouvelle édition du COG est intimement liée à l'organisation du recensement de population.***

Dès lors, la publication d'une nouvelle édition du COG se révèle intimement liée à l'organisation du recensement de la population. Il est en effet indispensable de disposer d'une base des communes exacte, afin d'organiser le recensement et de diffuser les populations de référence sur les bons territoires. Ce n'est pas un hasard si, parmi les treize éditions du COG parues jusqu'en 1999, sept d'entre elles sont publiées la même année qu'un recensement de la population (1954, 1961 en vue du recensement de 1962, 1968, 1975, 1982, 1990, 1999). Les évolutions territoriales ont cependant nécessité la publication de millésimes intercensitaires : en 1966 (pour tenir compte des vagues d'indépendance des anciens territoires français, en Afrique notamment), 1971, 1978 (avec l'apparition de nouveaux territoires d'outre-mer), 1985 (pour insérer les communes de Mayotte en vue de leur recensement) et 1994 (à la suite des bouleversements internationaux, notamment en Europe avec la fin du bloc de l'Est). Des rectificatifs sont également publiés régulièrement pour signaler les changements dans les objets du COG. Cette pratique, en vigueur dès les premières éditions du code, a été systématisée à partir de 1971 avec l'édition annuelle d'un livret qui répertorie les modifications territoriales intervenues au cours de l'année (et non plus des feuillets volants dactylographiés au rythme des changements).

Un événement majeur va bouleverser le rythme de publication du COG : la fin du recensement exhaustif et la mise en place de l'enquête annuelle de recensement. Depuis 2004, le recensement de la population a lieu tous les ans et les populations de référence sont diffusées annuellement. Il apparaît donc nécessaire de tenir le COG à jour sans prendre de retard sur les différents mouvements pouvant l'affecter. Ainsi, à partir des années 2000, le COG est actualisé à la date de référence du 1er janvier de chaque année. L'article 2 de l'arrêté ministériel du 28 novembre 2003 officialise ce rythme annuel de diffusion : « Le code officiel géographique est géré et publié par l'Institut national de la statistique et des études économiques (Insee) et mis à jour annuellement ».

21 Le COG constitue la partie centrale du référentiel des nomenclatures géographiques de l'Insee. Le champ de ce référentiel est bien plus vaste puisqu'il doit répondre aux besoins des statisticiens de l'institut. Il comprend évidemment des zonages d'études statistiques, comme les aires d'attraction des villes, mais aussi des zonages administratifs, tels que les établissements publics de coopération intercommunale (communautés de communes, communautés d'agglomération, etc.) ou encore électoraux (circonscriptions législatives). Il s'enrichit au fil du temps : par exemple, l'intégration de la nomenclature européenne des unités territoriales statistiques (NUTS) est actuellement à l'étude pour des besoins liés aux statistiques transfrontalières.

22 *https://www.insee.fr/fr/metadonnees/definition/c2297.*

23 *https://www.insee.fr/fr/metadonnees/definition/c2298.*

24 Voir les références juridiques en fin d'article.

▶ … et une référence pour les sphères publique et privée

> ***La référence commune à l'ensemble du service public français pour nommer et identifier des territoires.***

Outre le recensement de la population, le COG a acquis une visibilité et une notoriété qui ont contribué à asseoir son caractère de référence dans l'administration. Avec la loi du 7 octobre 2016 pour une République numérique et le décret du 14 mars 2017[25] relatif au service public de mise à disposition des données de référence, le COG devient la référence commune à l'ensemble du service public français pour nommer et identifier des territoires. Ce statut légal contribue à une large reprise de ses données par divers organismes, tant publics que privés. Elles entrent notamment dans la composition de nombreux identifiants administratifs construits à partir d'éléments géographiques. Ainsi, le code de la commune apparaît dans l'identité nationale de santé[26] gérée par l'Agence du numérique en santé (via le NIR), mais aussi dans les codes de traçabilité des entreprises agroalimentaires, ou encore dans le numéro des autorisations d'urbanisme comme le permis de construire. De même, le code du département apparaît dans le numéro FINESS des établissements sanitaires et sociaux.

Le COG nourrit des systèmes d'information essentiels pour l'action publique, tels que le système national de gestion des identifiants (SNGI), référentiel des identités pour les besoins des organismes de protection sociale (Préveraud de Vaumas, 2022). La gestion du SNGI requiert en effet de pouvoir établir une correspondance entre un code géographique et son libellé à une date donnée (la date de naissance). Le COG offre ainsi aux utilisateurs l'assurance d'une alimentation en données géographiques exactes et homogènes, assorties d'une profondeur historique. Ces propriétés en font une ressource fiable pour les systèmes d'information embarquant des services d'immatriculation ou d'identification d'individus.

▶ Un processus de production robuste…

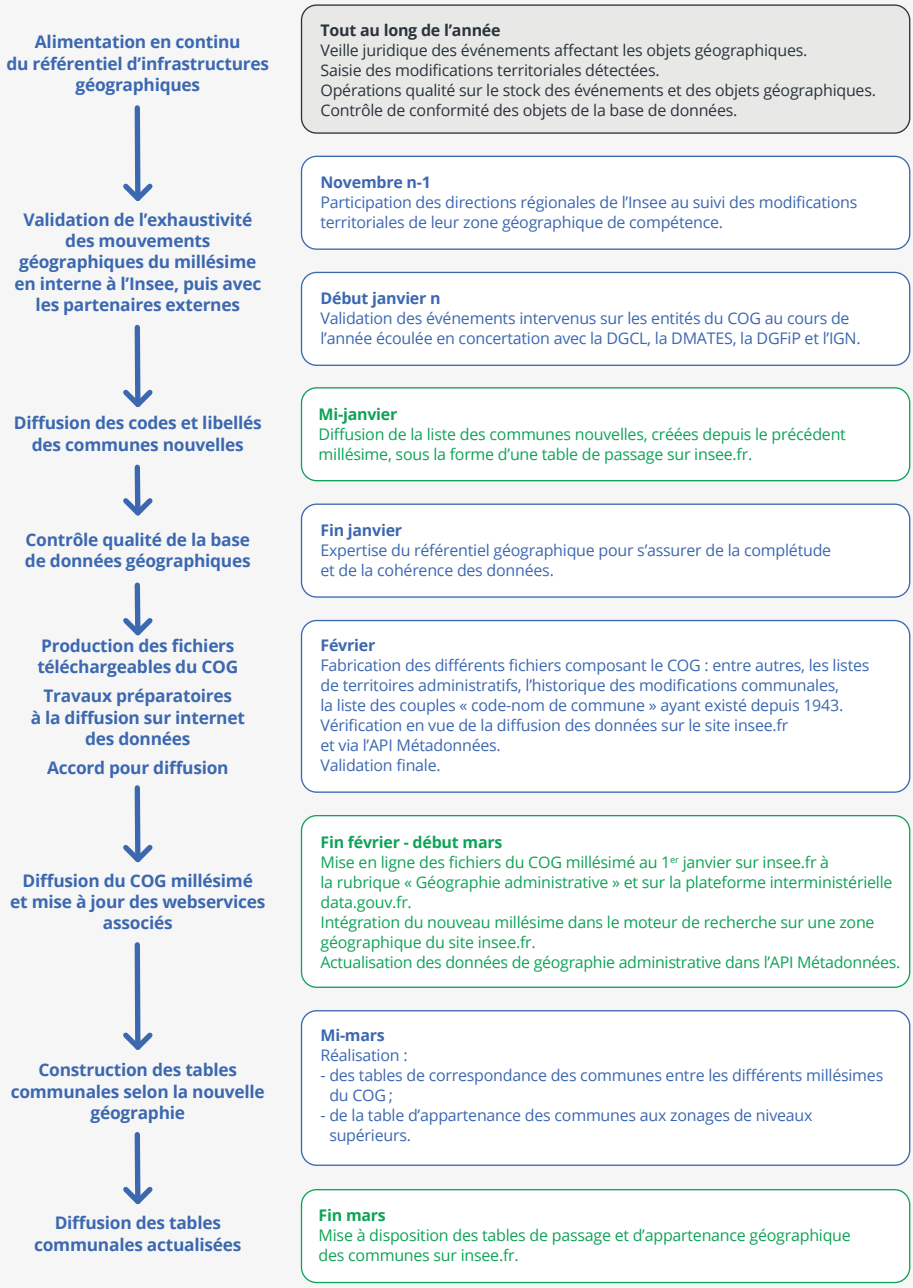
> ***La production repose sur une expérience acquise de longue date et une fréquence régulière d'actualisation.***

Le statut de jeu de données de référence implique de fortes exigences de qualité, les utilisateurs pouvant remonter leurs doléances par plusieurs canaux (messagerie, insee.fr, data.gouv.fr, etc.). La production du COG répond donc à des critères précis, entre autres d'exhaustivité, de fraîcheur, de cohérence, mais aussi, pour la mise à disposition, d'accessibilité, de format des données et de documentation.

Le processus est bien rodé et repose sur une expérience acquise de longue date et une fréquence régulière d'actualisation : l'Insee reprend dès 1946 la gestion et la publication de ce code statistique, auparavant assurées par le Service national des statistiques (« ancêtre » de l'Insee créé en 1941). Aujourd'hui, chaque millésime

25 Voir les références juridiques en fin d'article.
26 L'identité nationale de santé (INS) est un ensemble d'informations utilisé par les professionnels de santé pour identifier un patient dans son parcours de soin.

▶ Figure 3 - Les étapes de la production annuelle du COG : un calendrier resserré au 1er trimestre



API : *Application programming interface* (interface de programmation d'application).
DGCL : Direction générale des collectivités locales.
DMATES : Direction du management de l'administration territoriale et de l'encadrement supérieur.
DGFiP : Direction générale des Finances publiques.
IGN : Institut national de l'information géographique et forestière.

du COG est l'aboutissement d'une campagne annuelle qui se déroule en concertation avec un ensemble d'acteurs internes et externes à l'Insee, selon un calendrier contraint (***figure 3***). L'essentiel de l'information est obtenu par une veille réalisée tout au long de l'année par l'Insee et ses partenaires, permettant d'obtenir l'information au plus près de l'événement générateur (***figure 4***).

▶ … assis sur des textes administratifs et des partenariats multiples

En effet, tout changement géographique s'appuie sur un texte réglementaire qui décrit l'événement et précise les entités concernées. Cet acte prend la forme d'une délibération de conseil municipal, le plus souvent d'un arrêté préfectoral et parfois d'un décret ou d'une loi. Le Code général des collectivités territoriales contient les dispositions réglementaires encadrant les changements géographiques liés aux différentes collectivités territoriales (***encadré 1***).

L'Insee réalise donc une veille continue des textes publiés dans les Journaux officiels et les recueils des actes administratifs des préfectures départementales et régionales, mais aussi des projets de modifications territoriales relayés par la presse locale ou les bulletins municipaux. L'institut est de surcroît rapidement averti, par les préfectures de départements, des fusions et rétablissements de communes : ces dernières ont, comme les entreprises, besoin d'un numéro Siren[27] valide au répertoire Sirene pour leur activité courante.

En dehors de l'Insee, la démographie des communes est également suivie par des partenaires de premier plan. La direction générale des collectivités locales (DGCL) est naturellement très présente, au travers notamment du suivi de l'intercommunalité (***encadré 4***) et des communes nouvelles. La direction générale des finances publiques (DGFiP) suit les mouvements de création et de scission de communes, ceux-ci entraînant

▶ Encadré 4. Et les intercommunalités ? Une production du service statistique ministériel en charge des collectivités locales publiée en même temps que le COG

La réforme des collectivités territoriales de 2010 a obligé les communes à intégrer une intercommunalité à compter du 1er janvier 2014. À ce jour, toutes les communes appartiennent à un établissement public de coopération intercommunale (EPCI), à l'exception de quatre d'entre elles bénéficiant d'une dérogation : L'Île-d'Yeu, Île-de-Bréhat, Île-de-Sein et Ouessant.

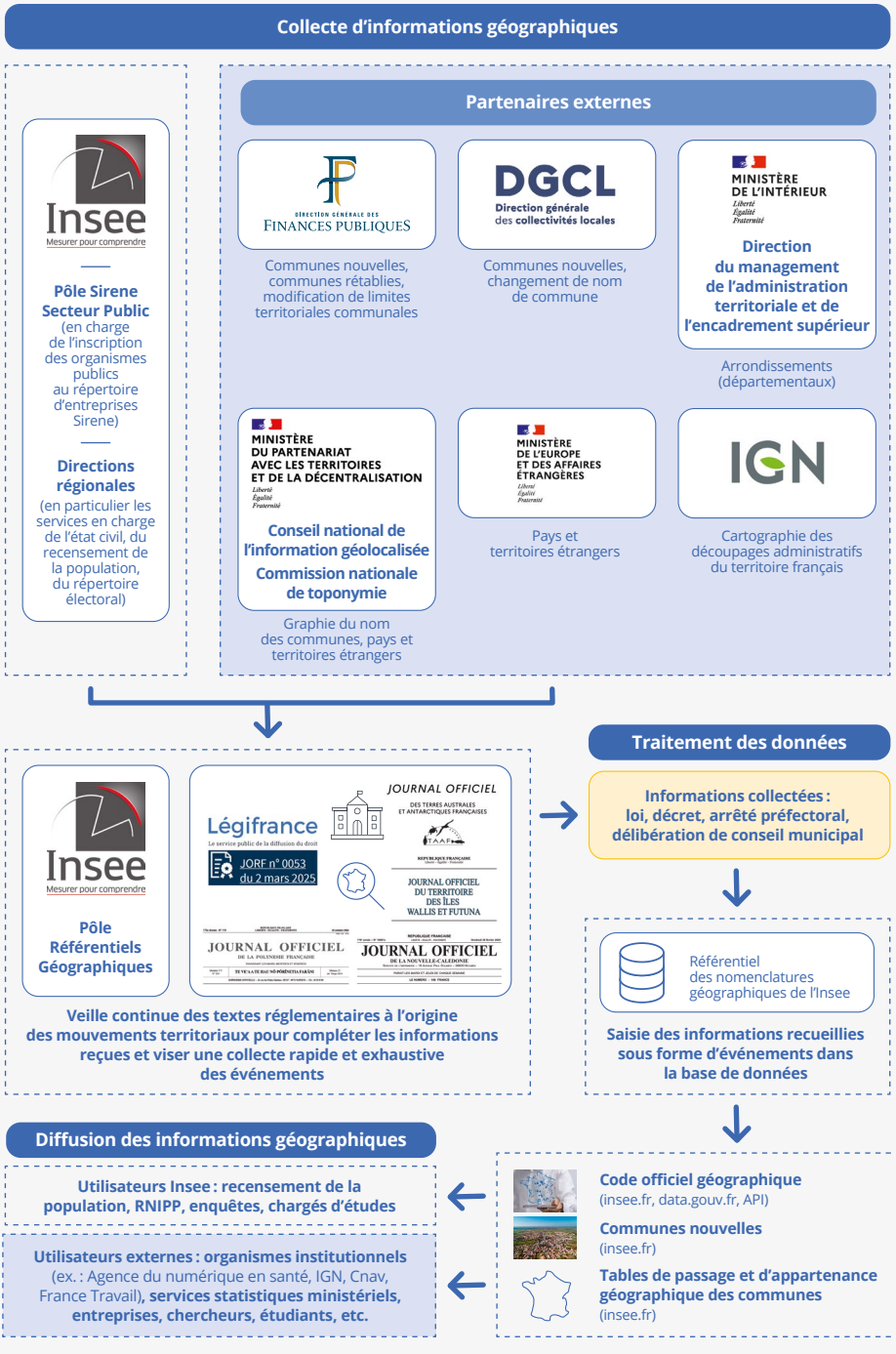
Les EPCI ne sont pas des collectivités territoriales au sens de l'article 72 de la Constitution. Toutefois, ils forment un zonage administratif dont le fonctionnement est régi par le Code général des collectivités territoriales (livre V).

L'Insee édite depuis 2012 une liste actualisée des EPCI à fiscalité propre* (communautés d'agglomération, communautés de communes, communautés urbaines, métropoles) en même temps que le COG. Néanmoins ce zonage n'est pas inclus dans la nomenclature, car sa propriété appartient à la direction générale des collectivités locales.

* *https://www.insee.fr/fr/information/2510634.*

27 *https://www.insee.fr/fr/metadonnees/definition/c1044.*

Figure 4 - Le circuit de l'information géographique : de la collecte à l'intégration dans le COG



IGN : Institut national de l'information géographique et forestière.
Cnav : Caisse nationale d'assurance vieillesse.

des conséquences fiscales et financières, ou encore les échanges de parcelles entre communes, pour la gestion du cadastre[28]. Enfin, l'IGN effectue une veille des évolutions des contours communaux pour produire les outils cartographiques les plus à jour.

▶ Un recueil non sans écueil

Toutefois, ce processus de recueil des informations géographiques n'est pas une sinécure. Plusieurs facteurs compromettent une collecte totale et rapide de tous les mouvements géographiques affectant le COG.

> ***Deux tiers des événements enregistrés dans le COG proviennent d'un arrêté préfectoral.***

D'une part, alors que les deux tiers des événements enregistrés dans le COG proviennent d'un arrêté préfectoral, la remontée d'informations des préfectures départementales vers l'Insee est assez aléatoire. Et si le COG est une compilation d'informations provenant de différents partenaires (***figure 4***), chaque partenaire suit une fraction limitée d'événements, certains champs n'étant que partiellement couverts (celui des communes déléguées, par exemple).

D'autre part, certaines modifications territoriales s'avèrent plus difficiles à repérer, en particulier celles entérinées par simple délibération de conseil municipal et ce, malgré de récentes mesures législatives pour favoriser la publicité numérique des actes pris par les communes.

C'est pourquoi des opérations de contrôle de la qualité du code sont réalisées. À titre d'exemple, en 2021 et 2022, une enquête menée auprès de communes créées par fusion a permis de supprimer environ 190 communes associées ou déléguées toujours présentes dans le COG.

La connaissance des directions régionales de l'Insee est aussi mise à contribution : elles sont ainsi sollicitées chaque année en novembre pour compléter l'inventaire des modifications territoriales du millésime en cours de constitution. En effet, de par leur implantation locale et la nature de leurs travaux, tels que le recensement de la population ou la gestion de l'état civil, elles entretiennent des relations de proximité avec les préfectures et les communes de leur ressort géographique.

▶ La validation du COG en point d'orgue

Une fois la collecte des informations effectuée, les parties prenantes se réunissent début janvier (***figure 3***) et valident conjointement l'ensemble des modifications apportées dans le COG, dont la liste des communes nouvelles diffusée vers le 15 janvier.

28 À noter que le cadastre fige sa géographie au 1er octobre N-1 pour la fiscalité à partir du 1er janvier N.

Toutes ces mises à jour sont intégrées dans le référentiel des nomenclatures géographiques de l'Insee par le biais d'événements (décrits par leur situation de départ et d'arrivée et les éléments d'information nécessaire). Ceci participe au suivi temporel de l'évolution des objets. Leur volumétrie varie souvent au gré des réformes territoriales (***figure 5***) : si par exemple la création du DROM de Mayotte a eu un impact limité, le régime des communes nouvelles instauré par la loi du 16 décembre 2010 de réforme des collectivités territoriales[29] et ses différentes adaptations ont eu des conséquences sur plusieurs années.

> ***La gestion du COG est régulièrement révisée pour s'adapter à une réglementation complexe et mouvante.***

La gestion du COG est régulièrement révisée pour s'adapter à une réglementation complexe et mouvante : l'exemple des collectivités territoriales à compétences départementales, hybrides entre une intercommunalité et un département, illustre l'inventivité de la loi. Cette phase, manuelle et technique, d'alimentation de la base de données conditionne in fine la qualité de restitution de l'information géographique aux utilisateurs et l'articulation entre les différents objets géographiques. C'est pourquoi l'Insee opère un ultime contrôle de la qualité des données, vérifiant leur exhaustivité et leur conformité, avant de produire les fichiers constitutifs du COG.

Après visa de la maîtrise d'ouvrage, le COG paraît généralement fin février, voire début mars, dans une version millésimée, photographie de la géographie administrative française et internationale au 1er janvier de l'année en cours.

▶ Une diffusion ancienne…

Pour nombre de personnes, la diffusion du code officiel géographique renvoie à un ouvrage épais, parfois poussiéreux, à l'instar d'autres ouvrages phares de l'institut comme la nomenclature d'activités française. On y trouve tout : les codes des départements, des arrondissements, des cantons, des communes, ainsi que des pays et territoires étrangers. Certains d'entre eux sont rares et précieux. Ainsi, il ne reste que peu d'exemplaires de la toute première édition du COG publiée en 1943, mais celle-ci nous donne une image de la France et du monde qui ne ressemble plus du tout au monde actuel : seulement trois départements en Île-de-France, les départements d'Algérie, les territoires français en Afrique et en Asie, etc. Les diffusions successives du COG témoignent des évolutions historiques importantes qu'ont connu la France et le monde ces quatre-vingts dernières années (Lang, 2003).

Aujourd'hui, on est très loin de cette image encyclopédique du COG. Les législations, les technologies et les besoins du public ont évolué. La diffusion du COG a dû (et su) s'adapter à ces changements, notamment depuis le début des années 2000 lorsque ce mouvement s'est accéléré.

29 Voir les références juridiques en fin d'article.

Une évolution importante est la gratuité, depuis 2003, de la diffusion des données du COG. Par exemple, l'édition 1999 coûtait 200 francs (30,49 euros) pour les utilisateurs externes alors que l'accès est désormais libre depuis le site insee.fr. Ce tournant du papier vers le numérique se fera progressivement : si la dernière édition papier du COG paraît en 1999, l'impression des livrets rectificatifs ne cesse qu'en 2018.

▶ Figure 5 - Les grands mouvements dans le COG en regard des événements législatifs depuis 15 ans

Édition 2012	**Les premières communes nouvelles font une entrée timide dans le COG** **2010 : loi de réforme des collectivités territoriales*** Naissance d'un nouveau dispositif de fusion de communes, la commune nouvelle, qui introduit le statut de commune déléguée. **Mayotte devient le 101e département français en mars 2011** **2010 : loi relative au département de Mayotte** Rattachement des 17 communes de l'archipel au département nouvellement créé.
Édition 2015	**Réforme de l'échelon cantonal** **2013 : loi entraînant une nouvelle délimitation des cantons fondée sur des critères démographiques** Leur nombre est réduit de moitié en février 2014. **Création de la métropole de Lyon, point de départ à l'introduction de la nomenclature des collectivités territoriales à compétences départementales** **2014 : loi de modernisation de l'action publique territoriale et d'affirmation des métropoles** Deux entités administratives coexistent désormais : le département en tant que circonscription administrative de l'État et la collectivité territoriale exerçant des compétences départementales.
Édition 2016	**Refonte de la carte régionale** **2015 : loi relative à la délimitation des régions** Le découpage passe de 22 à 13 régions en France métropolitaine, auxquelles s'ajoutent les 5 régions ultramarines. **Révision des limites territoriales des arrondissements départementaux** **2016 : circulaire sur la mise en œuvre de la réforme de l'échelon infradépartemental de l'État** Près de 1 900 communes changent d'arrondissement dans une soixantaine de départements.
Éditions 2016 à 2019	**Amélioration et adaptation du régime de la commune nouvelle** **2015 : loi relative à l'amélioration du régime de la commune nouvelle, pour des communes fortes et vivantes** **2016 : loi tendant à permettre le maintien des communes associées, sous forme de communes déléguées, en cas de création d'une commune nouvelle** **2019 : loi visant à adapter l'organisation des communes nouvelles à la diversité des territoires** En l'espace de 4 ans, près de 2 500 communes disparaissent, regroupées au sein de 793 communes nouvelles.
Éditions 2019 et 2020	**Ajustement du découpage cantonal au périmètre des communes nouvelles** Selon le Code général des collectivités territoriales, les communes de moins de 3 500 habitants sont entièrement incluses dans un même canton et non multicantonales, ce qui était jusqu'alors le cas d'une soixantaine de communes nouvelles.
Édition 2024	**Mamoudzou devient officiellement le chef-lieu de Mayotte après… 45 ans de statut provisoire !** **2023 : décret portant fixation du chef-lieu de Mayotte**

* Voir les références juridiques en fin d'article.

▶ … moderne pourtant…

> ***Les données du COG sont publiées numériquement depuis 2000 sur le site insee.fr.***

Les données du COG sont publiées numériquement depuis 2000 sur le site insee.fr. Elles ont un grand succès, directement par le biais de la rubrique « Géographie administrative et d'études » et indirectement par les données locales. Les élus, citoyens, étudiants, élèves et professeurs recherchent fréquemment les populations de référence, ou le dossier complet de chaque commune qui reprend l'ensemble des statistiques disponibles à ce niveau géographique.

Le COG a également bénéficié du mouvement d'ouverture des données : le développement de la *data* a favorisé la mise en place de référentiels partagés, produits par des acteurs reconnus comme l'Insee. La loi pour une République numérique de 2016 a ainsi marqué une étape dans la diffusion des données du COG : elle crée le service public de la donnée dont le COG est un des neuf jeux de données, publié depuis 2018 sur la plateforme interministérielle data.gouv.fr (Dinum, 2024).

Dernière évolution en date dans les modes de diffusion : depuis la fin de l'année 2020, les données du COG sont disponibles dans l'API Métadonnées[30] de l'Insee. Ce mode de diffusion, à l'adresse particulière des informaticiens, permet de mettre à jour facilement les systèmes d'information avec la dernière édition du COG. On trouve, pour chaque découpage, un service d'interrogation unitaire (pour un territoire donné) et un service de liste complète (pour un type de territoire), à une date donnée. D'autres services proposent un accès à l'historique du COG : prédécesseurs et successeurs de zonages (quelles communes ont fusionné pour former telle commune nouvelle), projections de zonages d'une date vers une autre (pour pouvoir reconstituer des périmètres comparables dans le temps). L'API comporte également des services permettant de naviguer dans la nomenclature des territoires et, par exemple, détailler les « contenants » et les « contenus » des zonages (communes d'un département, intercommunalités d'une région, etc.).

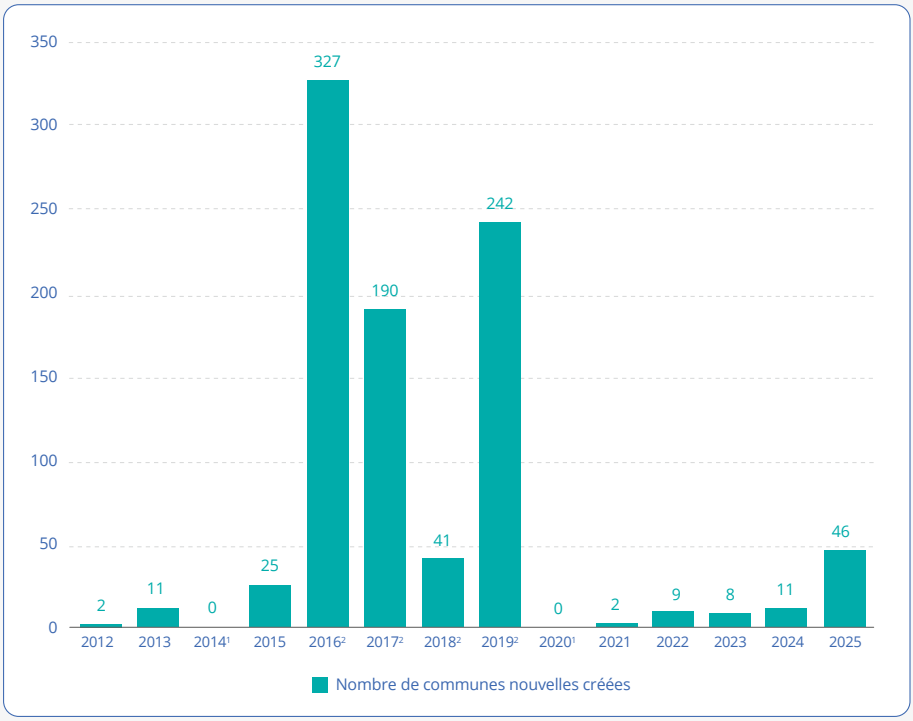
▶ … qui s'adapte à la loi et aux besoins des utilisateurs

Au cours de ces quinze dernières années, de nouvelles données ont été mises à disposition, consécutivement à des évolutions législatives ou des besoins.

Les communes déléguées sont diffusées avec le COG depuis le millésime 2015. De plus, à partir de 2016, un fichier spécifique des communes nouvelles créées l'année précédente (plus exactement entre le 2 janvier de l'année précédente et le 1er janvier de l'année en cours) est diffusé en avance de phase, aux alentours du 15 janvier. La production de ce fichier découle d'une très grande vague de créations de communes nouvelles issues de fusions d'anciennes communes, du fait des incitations financières introduites par la loi du 16 mars 2015 relative à l'amélioration du régime de la commune nouvelle[31]. Cette vague

30 *https://portail-api.insee.fr/catalog/api/5029cf12-e930-4d24-a9cf-12e9307d241d.*
31 Voir les références juridiques en fin d'article.

▶ Figure 6 - Création des communes nouvelles : un dynamisme sensible aux mesures incitatives



Notes :
1. Aucune commune nouvelle n'a vu le jour en 2014 et 2020 du fait de la tenue des élections municipales. En effet, le code électoral interdit le redécoupage des circonscriptions au cours de l'année précédant le premier tour du scrutin.
2. Sur la période 2016-2019, de nombreuses communes nouvelles issues de fusions d'anciennes communes ont été créées, après l'instauration de la loi du 16 mars 2015 relative à l'amélioration du régime de la commune nouvelle.

s'est poursuivie jusqu'en 2019 (***figure 6***) et la diffusion de ce fichier très demandé perdure aujourd'hui, même si le nombre de créations a diminué.

Depuis 2015, par ailleurs, l'Insee distingue les collectivités territoriales à statut particulier par une numérotation spécifique, en particulier celles ayant les compétences habituellement dévolues à un département. La première collectivité de ce type est la métropole de Lyon en 2015, qui se substitue sur son périmètre au département du Rhône. D'autres territoires ont enrichi cette liste au fil des années : le conseil départemental de Mayotte, la collectivité de Corse, les collectivités territoriales uniques de Martinique et de Guyane et, plus récemment, la Ville de Paris.

Depuis 2021, le COG dispose en outre d'un historique des couples code-commune remontant à 1943, afin de permettre aux utilisateurs de retrouver facilement la commune qui avait un code donné à une date particulière. Cette liste a été complétée en 2024 avec l'historique des couples code-pays, ainsi que la correspondance entre les codes

extension[32] et leur territoire de rattachement (commune ou pays) pour les besoins des répertoires d'individus. Dans la même veine, la correspondance entre une commune et ses différents zonages et ses prédécesseurs et successeurs est publiée sur la période 2003-2025. Il est donc possible de naviguer dans le temps et dans l'espace.

Enfin, l'édition 2022 a complété l'information sur les collectivités et territoires français d'outre-mer. Elle comprend un premier fichier avec leur liste et un second avec leurs zonages descendants : les communes de Saint-Martin, Saint-Barthélemy, Saint-Pierre-et-Miquelon, Nouvelle-Calédonie et Polynésie française ; les districts des Terres australes et antarctiques françaises ; les circonscriptions de Wallis-et-Futuna. L'historique des codes de ces territoires ultramarins, ainsi que celui des codes des communes d'Algérie, du Maroc et de Tunisie avant leur indépendance, ont été mis en ligne début 2025. Désormais, les utilisateurs disposent d'un panorama exhaustif des codes ayant existé dans le COG de 1943 à nos jours.

▶ Et ailleurs ?

Le besoin de décrire le territoire est ancien, mais il est également commun à tous les pays (ne serait-ce que pour la diffusion fine des recensements). Voici quelques points de comparaison sur les pratiques de nos voisins.

La Suisse dispose d'un répertoire officiel des communes (OFS, 2024) : les noms et les numéros de communes indiqués dans ce répertoire sont d'usage obligatoire pour les autorités depuis 2008, sachant que l'OFS attribue un numéro à chaque commune depuis 1960. En Italie, le code géographique est plus récent puisqu'il ne date que de 1991 sous le nom de *Codici statistici delle unità amministrative territoriali* (Istat, 2024). Mais il recense les mouvements territoriaux depuis 1861 et intègre les textes réglementaires associés. Le Portugal va plus loin encore : la première édition du COG portugais (*Código da divisão administrativa*) date de 1837 (INE, 2024).

Dernier exemple, le répertoire des communes allemand (*Gemeindeverzeichnis*), établi à des fins statistiques, comprend pour chaque commune, outre le code et le libellé, le code postal et les coordonnées géographiques du chef-lieu, la superficie, la population et la densité, ainsi que le degré d'urbanisation (Destatis, 2024). Particularité : il est mis à jour tous les trimestres.

Ces quelques exemples montrent la proximité des codes géographiques pour la statistique et donnent des pistes pour améliorer encore l'utilité du COG français.

32 Une utilisation des codes des communes, pays et territoires étrangers est l'inscription des individus au RNIPP. La gestion de cette base nécessite la création de codes extension pour enregistrer plus de 1 000 naissances sur un mois dans un territoire donné. Le 1 000e bébé (et chaque bébé suivant) aura dans son NIR non pas le code officiel géographique de sa commune (ou de son pays) de naissance, mais un code extension de celle-ci (ou de celui-ci).

▶ Fondements juridiques

- Constitution du 4 octobre 1958 : Titre XII : Des collectivités territoriales. In : *site de Légifrance*. [en ligne]. Mise à jour le 10 mars 2024. [Consulté le 6 janvier 2025]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/LEGISCTA000006095833*.

- Code général des collectivités territoriales. In : *site de Légifrance*. [en ligne]. Mise à jour le 7 novembre 2024. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006070633/*.

- Loi n° 71-588 du 16 juillet 1971 sur les fusions et regroupements de communes. In : *site de Légifrance*. [en ligne]. Mise à jour le 22 mars 2015. [Consulté le 6 janvier 2025]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/LEGITEXT000006068419*.

- Loi n° 2010-1563 du 16 décembre 2010 de réforme des collectivités territoriales. In : *site de Légifrance*. [en ligne]. Mise à jour le 1er janvier 2017. [Consulté le 6 janvier 2025]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000023239624*.

- Loi n° 2013-403 du 17 mai 2013 relative à l'élection des conseillers départementaux, des conseillers municipaux et des conseillers communautaires, et modifiant le calendrier électoral. In : *site de Légifrance*. [en ligne]. [Consulté le 6 janvier 2025]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000027414225*.

- Loi n° 2015-292 du 16 mars 2015 relative à l'amélioration du régime de la commune nouvelle, pour des communes fortes et vivantes. In : *site de Légifrance*. [en ligne]. [Consulté le 6 janvier 2025]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000030361485*.

- Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique (1). In : *site de Légifrance*. [en ligne]. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746*.

- Loi n° 2019-809 du 1er août 2019 visant à adapter l'organisation des communes nouvelles à la diversité des territoires. In : *site de Légifrance*. [en ligne]. Mise à jour le 3 août 2019. [Consulté le 6 janvier 2025]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000038864079/*.

- Loi n° 2022-217 du 21 février 2022 relative à la différenciation, la décentralisation, la déconcentration et portant diverses mesures de simplification de l'action publique locale (1) (dite loi 3DS). In : *site de Légifrance*. [en ligne]. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000045197395*.

- Décret n° 46-2948 du 30 décembre 1946 déclarant authentique les résultats du recensement de la population du 10 mars 1946. In : *site de Légifrance*. [en ligne]. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000685968*.

- Décret n° 2017-331 du 14 mars 2017 relatif au service public de mise à disposition des données de référence. In : *site de Légifrance*. [en ligne]. Mis à jour le 1er avril 2017. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000034194946#JORFTEXT000034194946*.

- Arrêté du 28 novembre 2003 relatif au code officiel géographique. In : *site de Légifrance*. [en ligne]. Mise à jour le 13 décembre 2003. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000798911*.

- Arrêté du 14 juin 2017 relatif aux règles techniques et d'organisation de mise à disposition des données de référence prévues à l'article L. 321-4 du code des relations entre le public et l'administration. In : *site de Légifrance*. [en ligne]. Mise à jour le 17 juin 2017. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000034944648*.

- Circulaire n° 131 du 21 août 1951 de la direction des routes modifiant la circulaire n° 48 du 11 mars 1950 sur l'immatriculation des véhicules automobiles. Abrogée par la circulaire du 20 juillet 1954. In : *site de Légifrance*. [en ligne]. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000506731*.

- Note d'information du 8 février 2021 de la direction générale des collectivités locales, relative à l'instruction de demandes de changement de nom des communes. In : *site cnig.gouv.fr*. [en ligne]. [Consulté le 3 avril 2025]. Disponible à l'adresse : https://cnig.gouv.fr/IMG/pdf/instruction-des-demandes-de-changement-de-nom-des-communes.pdf.

- Instruction ministérielle du 18 avril 2017 relative à la fixation du nom d'une commune nouvelle. In : *site cnig.gouv.fr*. [en ligne]. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://cnig.gouv.fr/IMG/documents_wordpress/2017/04/Indications_nom-commune-nouvelle_18042017.pdf*.

Bibliographie

- BENSOUSSAN, Johanna, BIZINGRE, Joël et COURVALIN, Nathalie, 2023. FINESS, le répertoire des établissements de santé. In : *Courrier des statistiques*. [en ligne]. 11 décembre 2023. Insee, N° N10, pp. 71-92. [Consulté le 11 décembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7722095?sommaire=7722116*.

- BIZINGRE, Joël, PAUMIER, Joseph et RIVIÈRE, Pascal, 2013. *Les référentiels du système d'information*. Juillet 2013. Dunod. Collection InfoPro. ISBN 978-2100598748.

- CARMILLE, René, 1942. *La mécanographie dans les administrations*. 2e édition, 1942, pp. 122-124. Recueil Sirey.

- COMMISSION NATIONALE DE TOPONYMIE (CNT), 2021. *Décider du nom d'un lieu. Guide pratique à l'usage des élus*. [en ligne]. Janvier 2021. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://cnig.gouv.fr/IMG/pdf/decider_du_nom_dun_lieu_01-2021.pdf*.

- DESTATIS (institut national de statistique allemand), 2024. Amtlicher Gemeindeschlüssel (AGS). In : *site de Destatis*. [en ligne]. [Consulté le 20 décembre 2024]. Disponible à l'adresse : *https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/Glossar/amtlicher-gemeindeschluessel.html*.

- DIRECTION INTERMINISTÉRIELLE DU NUMÉRIQUE, 2024. Service public de la donnée : des données sur lesquelles vous pouvez compter. In : *site gouvernemental des données ouvertes data.gouv.fr*. [en ligne]. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://www.data.gouv.fr/fr/pages/spd/reference/*.

- ESPINASSE, Lionel et ROUX, Valérie, 2022. Le Répertoire national d'identification des personnes physiques (RNIPP) au cœur de la vie administrative française. In : *Courrier des statistiques*. [en ligne]. 22 novembre 2022. Insee, N° N8, pp. 72-92. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665188?sommaire=6665196*.

- INE (institut national de statistique portugais), 2024. Código da divisão administrativa. In : *site de l'Ine*. [en ligne]. [Consulté le 20 décembre 2024]. Disponible à l'adresse : *https://smi.ine.pt/Versao*.

- INSEE, 1954. *Code officiel géographique 1954*.

- INSEE, SERVICES STATISTIQUES MINISTÉRIELS et INSTITUTS ET SERVICE DE STATISTIQUE TERRITORIAUX, 2023. Statistiques publiques dans les départements et régions d'outre-mer et les collectivités d'outre-mer. In : *Insee Méthodes*. [en ligne]. 21 septembre 2023. Insee, N° 144, pp. 1118. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7641151*.

- INSTITUT NATIONAL DE L'INFORMATION GÉOGRAPHIQUE ET FORESTIÈRE (IGN), 2024. ADMIN EXPRESS. Le découpage administratif du territoire français. In : *site géoservices de l'IGN*. [en ligne]. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://geoservices.ign.fr/adminexpress*.

- ISTAT (institut national de statistique italien), 2024. SITUAS : il nuovo portale Istat per la gestione del territorio. In : *site de l'Istat*. [en ligne]. [Consulté le 11 décembre 2024]. Disponible à l'adresse : *https://www.istat.it/wp-content/uploads/2024/04/Situas.pdf*.

- LANG, Gérard, 2003. Le Code officiel géographique. In : *Courrier des statistiques*. [en ligne]. Décembre 2003. Insee, N° 108, pp. 53-62. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://www.bnsp.insee.fr/ark:/12148/bc6p06xt2c8/f1.pdf*.

- LANG, Gérard, 2017. *Le code officiel géographique (COG), avant, pendant et autou*r (notes, annexes et tableaux). [en ligne]. Janvier 2017. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *http://www.christophe-terrier.fr/CT/Textes/gl2017_COG-2017-01-04-continu.pdf* ***(note principale),*** *http://www.christophe-terrier.fr/CT/Textes/gl2017_COG-2017-01-04-annexes-et-tableaux.pdf* ***(note complémentaire).***

- MAUGUIN, Jocelyne et SAGNES Nicolas, 2024. Faciliter l'accès aux données de l'Insee – Cubes, catalogue et métadonnées. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2024. Insee. N° N11, pp. 31-50. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8203036?sommaire=8203072*.

- OFFICE FÉDÉRAL DE LA STATISTIQUE (OFS), 2024. Répertoire officiel des communes de Suisse. In : *site de l'OFS*. [en ligne]. [Consulté le 11 décembre 2024]. Disponible à l'adresse : *https://www.bfs.admin.ch/bfs/fr/home/bases-statistiques/repertoire-officiel-communes-suisse.html*.

- PRÉVERAUD DE VAUMAS, Joseph, 2022. Un référentiel des identités pour les besoins de la sphère sociale. Le système national de gestion des identifiants (SNGI). In : *Courrier des statistiques*. [en ligne]. 29 novembre 2022. Insee, N° N8, pp. 93-114. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665190?sommaire=6665196*.

- RIVIÈRE, Pascal, 2022. Qu'est-ce qu'un répertoire ? De multiples exigences pour un système complexe. In : *Courrier des statistiques*. [en ligne]. 29 novembre 2022. Insee. N° N8, pp. 52-71. [Consulté le 7 novembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665186?sommaire=6665196*.

Sources de données privées : panorama et perspectives

*Romain Lesur**

Depuis la fin du XXe siècle, le numérique a transformé l'économie, générant massivement des traces numériques exploitables par les entreprises et désormais par le système statistique public. Ce dernier s'intéresse à ces données pour compléter les enquêtes traditionnelles, recherchant une plus grande granularité spatiotemporelle, tout en tenant compte des enjeux technologiques, juridiques et méthodologiques qui leur sont associés.

Initialement appelées « big data », ces traces numériques sont aujourd'hui désignées comme des « données détenues par des opérateurs privés », marquant un déplacement d'enjeux vers les questions d'accès et de qualité. L'Insee a intégré les données de caisse des grandes surfaces alimentaires dans la production de l'indice des prix à la consommation et a étendu ses travaux exploratoires aux données de téléphonie mobile, de transactions par carte bancaire, de comptes bancaires ou encore de locations touristiques de courte durée.

Ces nouvelles sources enrichissent la connaissance économique et sociale, mais présentent des limites : couverture partielle, imprécisions géographiques ou temporelles, et absence de variables sociodémographiques. Leur exploitation nécessite des partenariats renforcés avec des opérateurs privés, des infrastructures sécurisées, et une méthodologie rigoureuse.

La stratégie européenne pour les données et la révision du règlement européen no 223/2009 offrent désormais un cadre pour faciliter l'accès à ces sources. Malgré leur potentiel, ces dernières restent souvent complémentaires aux enquêtes, chacune ayant ses forces et faiblesses, rendant indispensable leur articulation méthodologique afin d'enrichir les statistiques publiques.

Since the end of the 20th century, digital technology has transformed the economy, generating data at scale that can be exploited by businesses and now by the official statistical system. The latter is interested in this data to supplement traditional surveys, seeking greater spatiotemporal granularity, while taking into account the technological, legal and methodological issues associated with these new data sources.

Initially referred to as 'big data', these data are now referred to as 'privately held data', marking a shift in the focus towards data access and quality. INSEE has integrated scanner data from supermarkets into its consumer price index, and has extended its exploratory work to mobile network operators data, payment bankcards transactions, bank accounts and short-term tourist accommodation.

These new sources enrich our economic and social knowledge, but have their limitations: partial coverage, geographical or temporal inaccuracies, and the absence of socio-demographic variables. Their use requires stronger partnerships with private operators, secure infrastructures and a rigorous methodology.

The European data strategy and the revision of Regulation No 223/2009 now provide a framework to facilitate access to these sources. Despite their potential, these sources are still often complementary to surveys, each with its own strengths and weaknesses, making it essential to combine them methodologically in order to enrich official statistics.

* Chef de l'unité SSP Lab, Insee.
romain.lesur@insee.fr

Depuis la fin du XXe siècle, l'économie s'est transformée sous l'effet du développement du numérique (Insee, 2019). Les entreprises ont modernisé leur fonctionnement, tant en interne, avec par exemple le recours fréquent à des systèmes intégrés de gestion[1], que dans leurs relations externes, avec notamment l'essor du commerce électronique. Conséquence de ces évolutions, **de très nombreux pans de l'activité des individus et des entreprises donnent désormais lieu à des traces numériques**.

Ces nouvelles données détenues par les entreprises représentent une opportunité pour celles-ci, car elles leur permettent d'optimiser leurs processus internes ou d'améliorer les services offerts à leurs clients (Vacher et Pradines, 2017). Depuis une quinzaine d'années, le système statistique public s'y intéresse également. Il y voit en effet un moyen de compléter ses sources d'information traditionnelles, à savoir les grandes enquêtes qu'il réalise, qui restent encore la fondation du système. Il cherche en particulier à bénéficier d'une plus grande granularité spatiotemporelle, tout en tenant compte des limites méthodologiques de ces données liées à des processus de gestion (Blanchet et Givord, 2017).

Où en est-on aujourd'hui de l'utilisation de ces traces numériques par le système statistique public ?

▶ Des « big data » aux « données détenues par des opérateurs privés »

Durant les années 2010, le terme le plus communément employé pour désigner ces sources était celui de **données massives** (ou **big data**). Le choix d'un tel vocabulaire soulignait en premier lieu le défi technologique que représentaient leur stockage et leur traitement. Les progrès réalisés depuis une quinzaine d'années ont désormais permis de résoudre ces enjeux, à tel point que le terme « big data » est de moins en moins usité (Tigani, 2023). La statistique publique maîtrise aujourd'hui ces technologies, qui constituent les soubassements des nouvelles plateformes de *data science* telles que le SSP Cloud (Comte et al., 2022). Ces infrastructures facilitent le traitement de données massives, notamment par le recours à de nouveaux formats de données (Dondon et Lamarche, 2023).

Désormais, ces traces numériques sont désignées sous le nom de **données détenues par des opérateurs privés**. Le glissement sémantique n'est pas anodin : parmi les multiples défis qu'elles posent, celui de la technologie aura été certainement le plus vite résolu. Ces nouvelles sources représentent donc le troisième type de données traité par la statistique publique, aux côtés des enquêtes et des données administratives. Elles partagent avec les données administratives le fait qu'elles n'ont pas été produites à des fins de statistique publique et qu'elles doivent donc d'abord être « qualifiées »[2], avant d'être éventuellement intégrées dans le système d'information statistique (Cotton et Haag, 2023).

1 Systèmes permettant de centraliser et de rationaliser l'ensemble des données de gestion des entreprises (ressources humaines, comptabilité, activité commerciale, etc.).

2 Selon Cotton et Haag (2023), « il est nécessaire d'échanger avec le producteur de la donnée afin de vérifier que la source est : exploitable (les données contenues peuvent être restructurées pour mesurer des concepts statistiques) ; complète (aucune sous-couverture évidente qui empêcherait son exploitation) ; disponible dans un délai raisonnable ; documentée (présence de métadonnées) ».

> ***La première source de données détenues par des opérateurs privés exploitée par l'Insee dans le cadre de sa production statistique correspond aux données de caisse des enseignes de la grande distribution alimentaire.***

La première source de données détenues par des opérateurs privés exploitée par l'Insee dans le cadre de sa production statistique correspond aux données de caisse des enseignes de la grande distribution alimentaire. L'Insee les exploite depuis 2020 pour élaborer l'indice des prix à la consommation (Leclair, 2019). À ce jour, il s'agit de la principale source de ce type pleinement intégrée dans un processus de production statistique de l'Insee.

La crise sanitaire de 2020 a renforcé des collaborations déjà établies et en a suscité de nouvelles, avec des détenteurs de données soucieux de leur responsabilité sociétale. Ces acteurs ont souhaité mettre leurs informations à disposition des pouvoirs publics pour les aider à agir durant cette situation d'urgence. Certaines collaborations se sont prolongées à l'issue de la crise sanitaire. L'Insee, et plus largement les systèmes statistiques français et européen, ont ainsi pu approfondir l'examen de ces nouvelles sources afin de qualifier leurs potentiels. Les travaux se sont démultipliés : ils explorent l'utilisation des données détenues par des banques, des plateformes de location immobilière de courte durée, des opérateurs de téléphonie mobile, mais aussi celles des programmes de fidélité (Galiana et Suarez Castillo, 2022), des compteurs communicants (Le Saout et al., 2024) ou encore des « schémas » de paiement par carte bancaire (voir ci-dessous, dans la partie consacrée aux données de transactions par carte bancaire)[3].

Sur le long terme, et dans l'optique de leur intégration dans le système d'information statistique, les sources de données privées présentent **trois grands défis pour le service statistique public** (***figure 1***). Le premier est celui de la base légale sur laquelle il pourra se fonder pour y accéder dans le cadre de ses missions. Le deuxième porte sur la mise en place d'un cadre partenarial soutenable et pérenne pour traiter ces informations sans rompre leur confidentialité. Enfin, le troisième défi est d'ordre méthodologique, ces données n'ayant pas été collectées pour la réalisation de statistiques publiques.

Les nouvelles sources de données examinées par la statistique publique

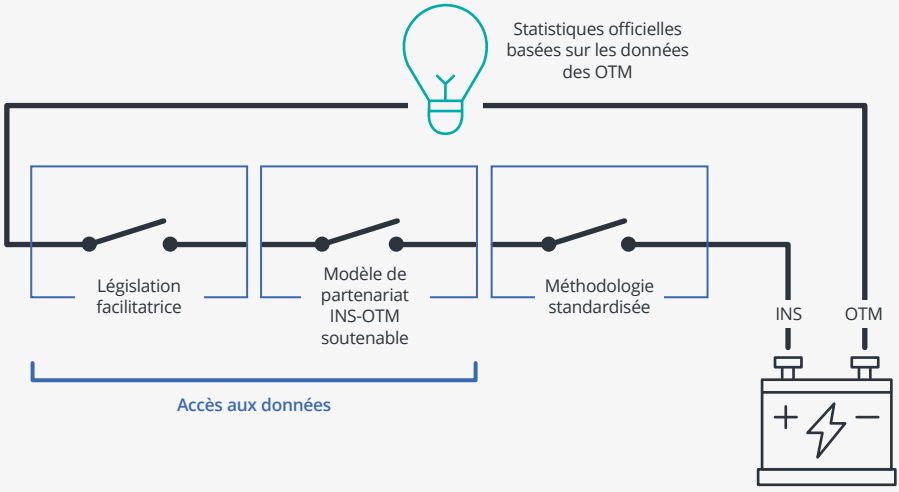
Après les données de caisse, l'Insee s'est donc intéressé à de nouvelles sources de données détenues par les opérateurs privés. Cet article en donne ici un panorama général (***figure 2***). De manière plus détaillée, les données de comptes bancaires ont fait l'objet d'un article dans le numéro précédent du *Courrier des statistiques* (Bonnet et Loisel, 2024). Les données de téléphonie mobile et de transactions par carte bancaire font, elles, l'objet des deux articles suivants de ce numéro[4].

3 L'Insee a également exploré le potentiel des données recueillies par Google au travers du service Google Trends pour compléter les enseignements de la statistique publique. L'apport de ces données est cependant limité (Bortoli et Combes, 2015 ; Tavernier et Ourliac, 2020).

4 Voir les articles de Joubert sur les données de téléphonie mobile et de Boittelle et al. sur les données de transactions par carte bancaire CB dans ce même numéro.

▶ Figure 1 - « Fermer le circuit »*

« La production régulière de statistiques officielles basées sur les données des opérateurs de téléphonie mobile ne pourra être mise en place que lorsque toutes les questions en suspens auront été résolues. La définition d'une méthodologie appropriée est une question logiquement distincte de la définition d'un modèle durable d'accès aux données, qui implique l'établissement de modèles de partenariats soutenables entre l'institut national de statistique et les opérateurs de téléphonie mobile dans le cadre d'une législation facilitatrice. »*



INS : Institut national de statistique.
OTM : Opérateur de téléphonie mobile.
* D'après la figure 1 de l'exposé de position d'Eurostat sur l'utilisation à des fins de statistique publique des données de téléphonie mobile (Eurostat, 2023).

Les données de téléphonie mobile

Les données de téléphonie mobile offrent un potentiel désormais reconnu pour compléter la connaissance statistique sur la population et les territoires. Elles permettent d'analyser la présence et la mobilité des personnes dans une zone géographique. Ainsi, il est possible d'estimer la population présente à différents moments de la journée, de la semaine ou de l'année, de mesurer les déplacements quotidiens et même de créer des cartographies dynamiques de la population. Ces données fournissent également des éclairages sur les flux touristiques et la ségrégation résidentielle (Galiana et al., 2020).

L'exploitation de ces données présente **plusieurs avantages indéniables**. Leur richesse informationnelle, leur fréquence d'actualisation et leur granularité géographique complètent les sources traditionnelles de la statistique publique, en tout premier lieu le recensement de la population. Elles peuvent permettre d'analyser finement les déplacements domicile-travail, et par exemple d'appréhender les effets du télétravail, mais aussi de comparer la mobilité entre la semaine et le week-end. Elles peuvent également apporter un éclairage utile aux acteurs publics locaux sur les besoins de protection de la population (services de santé et d'urgence, sécurité et protection civile) ou aux entreprises du commerce de détail sur les besoins en points de vente.

> ***Les données de téléphonie mobile permettent d'analyser finement les déplacements domicile-travail, d'appréhender les effets du télétravail, mais aussi de comparer la mobilité entre la semaine et le week-end.***

Cependant, leur utilisation soulève **de nombreux enjeux méthodologiques**. Les données de téléphonie mobile sont conçues et optimisées pour la gestion d'un réseau de télécommunications, et non pour la production de statistiques. Leur réutilisation à cette fin nécessite des traitements complexes et des méthodologies spécifiques (Suarez Castillo et al., 2023). Les principales difficultés sont les suivantes :

- L'enregistrement des interactions entre les téléphones mobiles et les antennes relais est irrégulier, en fonction par exemple de l'activité du téléphone, de la densité locale des antennes et du nombre d'utilisateurs qui y sont connectés. Cela engendre une incertitude temporelle pour l'estimation du nombre de personnes présentes dans une zone géographique, et ce tout particulièrement la nuit, lorsque des téléphones sont éteints ou en mode avion.

- La localisation des téléphones mobiles dépend de la couverture des antennes relais de chaque opérateur sur le territoire. Elle est approximative et plus ou moins précise selon les zones géographiques, en fonction de la densité de ces antennes. Cela crée une incertitude spatiale, qui peut aussi être plus élevée la nuit lorsque certaines antennes relais sont éteintes.

- La représentativité des données est partielle, si l'on se limite à un seul opérateur ; chacun ne couvre en effet qu'une partie de la population (Sakarovitch et al., 2019). L'estimation du nombre de personnes présentes à partir des données de téléphonie mobile se heurte aussi au fait que certaines personnes peuvent posséder plusieurs appareils ou abonnements, potentiellement chez différents opérateurs, ou inversement que le même appareil peut être utilisé par plusieurs personnes différentes.

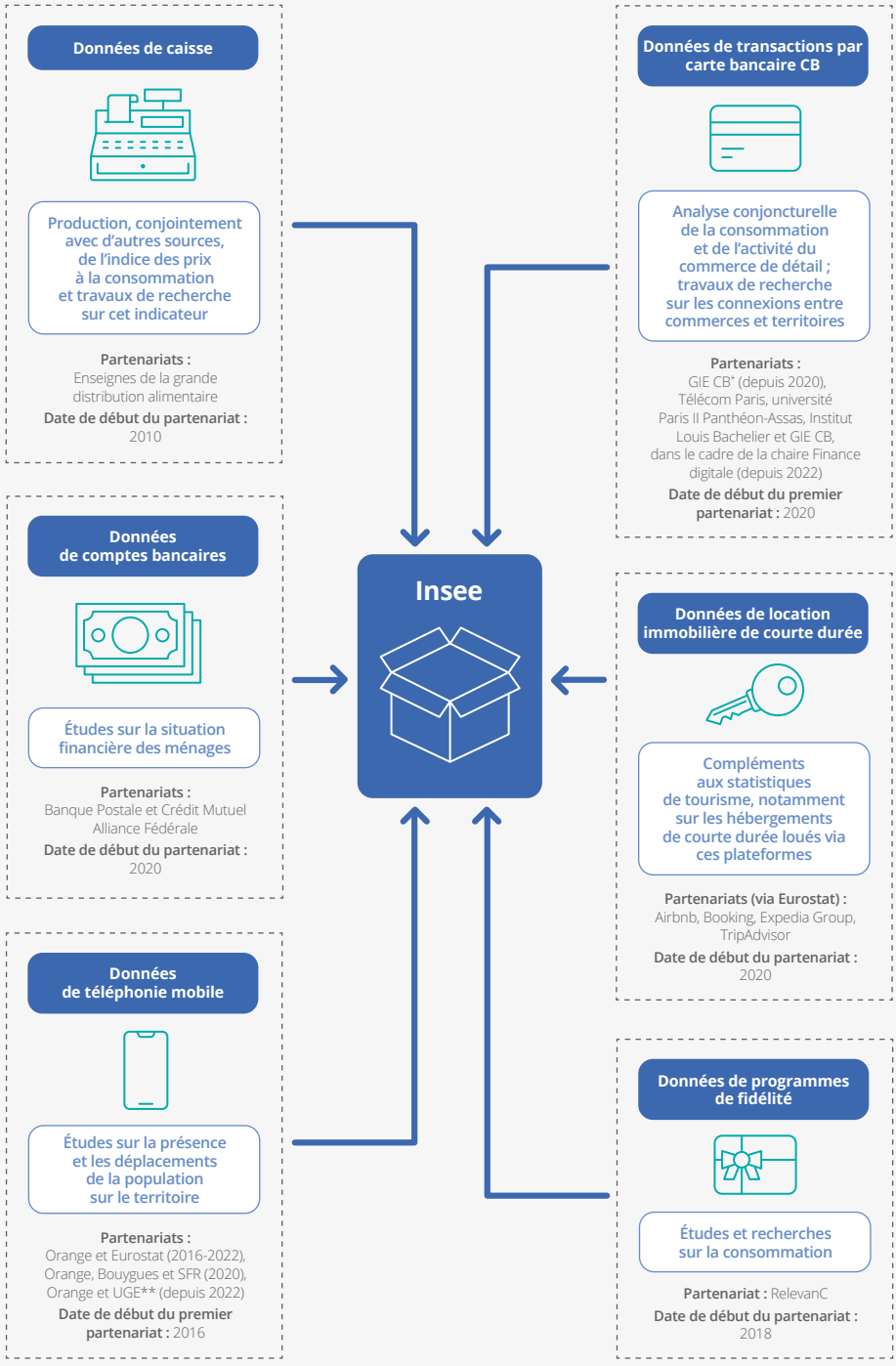
- L'absence de variables sociodémographiques précises concernant les utilisateurs d'appareils mobiles limite les possibilités d'exploitation à des fins de statistique publique. L'hétérogénéité des formats de données et des informations chez les différents opérateurs requiert par ailleurs des traitements complexes.

- Il est indispensable également de définir clairement les concepts mesurés, par exemple celui de « touriste », qui est parmi les plus complexes à aborder à partir des données de téléphonie mobile. En effet, un client d'un opérateur étranger en itinérance peut être détecté sur plusieurs réseaux mobiles français, conduisant ainsi à de multiples comptages.

- Enfin, la stabilité des indicateurs peut être affectée par l'évolution rapide des usages et des technologies.

L'utilisation des données de téléphonie mobile soulève également **des enjeux juridiques importants**. Les questions de confidentialité, de protection de la vie privée et d'accès aux données sont centrales s'agissant de données aussi sensibles que celles de téléphonie mobile. La protection de la vie privée, le respect du règlement général sur la protection des données (RGPD) et la protection du secret des affaires doivent être garantis à chaque étape du processus de traitement de l'information. Ceci impose une organisation et des investissements spécifiques.

▶ Figure 2 - Les données d'opérateurs privés déjà explorées par l'Insee



* GIE CB : Groupement d'intérêt économique Cartes Bancaires CB.
** UGE : Université Gustave Eiffel.

L'exploitation des données de téléphonie mobile nécessite donc une compréhension approfondie de leurs caractéristiques et limites, l'adoption de méthodologies adaptées et la mise en place d'un cadre technique, organisationnel et juridique qui réponde aux exigences et obligations de l'ensemble des acteurs concernés (Coudin et al., 2021[5]). Pour ces raisons, Eurostat[6] a lancé différents projets sur ces données, dont deux aboutiront en 2025 (***encadré 1***) :

- Le premier projet, intitulé « Multi-MNO », offrira une implémentation de référence d'un processus de traitement des données de téléphonie mobile garantissant leur confidentialité, depuis les opérateurs de téléphonie mobile vers les instituts nationaux de statistique.

- Le second projet, intitulé « MNO-MINDS », proposera un cadre méthodologique de référence pour combiner les données de téléphonie mobile avec d'autres sources, au sein d'un processus de production statistique cohérent.

▶ Encadré 1. Les projets européens sur les données de téléphonie mobile en 2025

En 2023, Eurostat a attribué des financements européens à deux projets portant sur les données de téléphonie mobile et dans lesquels l'Insee est impliqué : Multi-MNO et MNO-MINDS.

Le projet Multi-MNO* :

Ce projet, auquel l'Insee et Orange France sont associés dans le cadre de son comité consultatif, vise à définir, à des fins de production statistique, un ensemble de traitements standardisés de données ou *pipeline* (mis en œuvre chez les opérateurs de téléphonie mobile) et de flux de données (vers les instituts nationaux de statistique européens). Les traitements et flux de données ont été définis par le système statistique européen en partenariat avec les opérateurs. Afin à la fois de faciliter le déploiement du *pipeline* et la transparence des méthodologies appliquées chez les opérateurs, le projet propose une implémentation open source. Le *pipeline* a pour caractéristique de protéger intégralement la confidentialité : seules des données agrégées seront transmises aux instituts de statistique, données qui auront donc été calculées suivant des méthodologies communes et transparentes. Ce projet vise également à démontrer la faisabilité de mise en œuvre du *pipeline* en l'appliquant à des données réelles.

Le projet MNO-MINDS :**

Ce projet est coordonné par l'institut national de statistique italien (Istat) et associe dix partenaires, dont l'Insee. L'objectif est de proposer des méthodes, ainsi que des librairies open source, dédiées à l'intégration des données de téléphonie mobile avec d'autres sources, à des fins de production régulière de statistiques publiques. Ce projet proposera également des formations à ces méthodes et outils.

* *https://cros.ec.europa.eu/landing-page/multi-mno-project.*
** *https://cros.ec.europa.eu/mno-minds.*

Les données de transactions par carte bancaire

Les données issues des transactions par carte bancaire apportent des informations complémentaires à celles de la statistique publique, notamment pour le **suivi conjoncturel et la prévision des indicateurs économiques**. Une transaction par carte bancaire mobilise sept acteurs : l'acheteur, le commerçant, leurs banques respectives, le schéma de paiement (CB, Visa, Mastercard, etc.), le réseau interbancaire d'autorisation et le système interbancaire d'autorisation[7]. L'ensemble des transactions génère de

5 Voir également l'article de Joubert sur les données de téléphonie mobile dans ce même numéro.
6 Eurostat est l'institut statistique communautaire, direction générale de la Commission européenne.
7 Voir l'article de Boittelle et al. sur les données de transactions par carte bancaire CB dans ce même numéro.

> ***Les données de transactions par carte bancaire peuvent améliorer le suivi conjoncturel et la prévision de certains indicateurs d'activité, comme l'indice de chiffre d'affaires du commerce de détail.***

nombreuses données, dont les montants, dates et heures des opérations et les références bancaires des acheteurs et vendeurs. Une fois agrégées et enrichies, par exemple avec l'activité principale et la localisation du commerçant, ces données permettent d'analyser en profondeur les comportements de consommation et l'activité commerciale.

Face aux délais de collecte et de traitement de ses sources de données traditionnelles, l'Insee a cherché à mobiliser ces données, à haute fréquence et disponibles rapidement. Elles peuvent en effet améliorer le suivi conjoncturel et la prévision de certains indicateurs d'activité, comme l'indice de chiffre d'affaires du commerce de détail. Les données du groupement d'intérêt économique Cartes Bancaires CB (GIE CB)[8], qui est le schéma domestique français de paiement par carte et par mobile, couvrent un large spectre d'activités commerciales. Elles offrent un aperçu précis des dynamiques territoriales et sectorielles et ont notamment permis d'éclairer les conséquences économiques du confinement de 2020 (Insee, 2020).

Le partenariat entre l'Insee et le GIE CB – renforcé depuis la crise sanitaire et via la chaire de recherche Finance digitale[9] – repose sur une transmission sécurisée à l'Insee de données préagrégées. Ces dernières sont issues d'un travail de structuration et d'enrichissement des données individuelles anonymisées. Le dispositif permet l'exploitation d'informations détaillées sur chaque transaction tout en garantissant leur confidentialité. Des contrôles rigoureux, basés sur la comparaison entre données d'autorisation, de compensation et d'activité de paiement, permettent d'analyser la qualité et la fiabilité des agrégats exploités pour la production statistique.

Ces données présentent cependant **des limites intrinsèques**, qui nécessitent des corrections pour ne pas fausser les analyses :

- Elles ne couvrent qu'une partie des transactions réalisées en France. Elles ne concernent en effet que les paiements effectués par des personnes résidant en France auprès de commerçants affiliés à une banque française, et uniquement par carte bancaire CB. Elles n'intègrent donc pas les paiements en espèces ou par chèque, les virements ou les paiements en ligne via d'autres schémas de paiement que CB (par exemple les schémas internationaux Visa et Mastercard).
- Elles ne permettent pas de distinguer, au sein des paiements par carte, les dépenses professionnelles des dépenses personnelles.
- Elles sont sensibles à l'évolution de la couverture des dépenses par le schéma de paiement CB, liée notamment à la concurrence entre schémas et aux changements éventuels de comportements (par exemple l'essor du paiement par mobile).

8 *https://www.cartes-bancaires.com/cb/groupement/.*
9 *https://digital-finances.com/.* Voir à ce sujet l'article de Boittelle et al. sur les données de transactions par carte bancaire CB dans ce même numéro.

- Enfin, des imprécisions sur l'activité et la localisation des commerçants peuvent altérer l'analyse.

Durant la crise sanitaire, le caractère inframensuel de ces données et leur rapidité de mise à disposition ont permis d'obtenir des informations précieuses pour l'analyse conjoncturelle. Néanmoins, leur utilisation sur le moyen terme pour cette finalité reste encore une question ouverte, en raison d'une forte volatilité et d'une couverture fluctuante. En revanche, le chaînage des transactions pour une même carte offre la possibilité d'éclairer les **liens entre commerces et territoires** (c'est-à-dire où vont les habitants d'un territoire pour faire leurs achats). Il permet d'envisager l'utilisation de ces données pour analyser les effets de l'implantation ou de la disparition de certains points de vente sur l'activité commerciale infracommunale, ou bien l'évaluation de politiques publiques portant sur l'activité commerciale. Les comportements effectifs de consommation de la population observés à travers ces données peuvent également compléter les sources d'information déjà existantes sur les connexions entre territoires. En effet, les zonages actuels élaborés par l'Insee établissent ces connexions principalement à partir des déplacements domicile-travail (aires d'attraction des villes[10]) ou des distances des populations aux équipements les plus proches (bassins de vie[11]).

Au total, ces données enrichissent la palette des indicateurs économiques en offrant une vision plus fine des comportements de consommation et des dynamiques territoriales, tout en posant des défis méthodologiques pour leur intégration dans les processus statistiques traditionnels.

Les données de comptes bancaires

Le numéro précédent du Courrier des statistiques présente les travaux menés par l'Insee à partir d'échantillons anonymisés de comptes bancaires de La Banque Postale et du Crédit Mutuel Alliance Fédérale (Bonnet et Loisel, 2024). L'institut s'est intéressé à ces données en tout premier lieu pour l'**analyse de la conjoncture économique**. En effet, leur fraîcheur, leur granularité fine, la grande taille des échantillons considérés et la variété des renseignements disponibles favorisent des analyses précises et quasiment en temps réel des comportements financiers des ménages. Ainsi, les relevés mensuels de comptes bancaires, délivrés dès la fin du mois suivant la période observée, voire au milieu du mois, permettent d'étudier les revenus et dépenses des ménages sur une base journalière. La diversité des informations recueillies autorise également une analyse fine de la **diversité des comportements des ménages** et des **inégalités de consommation** (y compris sur des trajectoires longues, grâce à la possibilité de constituer des panels). Des populations spécifiques peuvent par ailleurs être étudiées avec davantage de précision, car les effectifs disponibles sont plus importants que dans les enquêtes traditionnelles.

> ***Les relevés mensuels de comptes bancaires permettent d'étudier les revenus et dépenses des ménages sur une base journalière.***

10 *https://www.insee.fr/fr/metadonnees/definition/c2173.*
11 *https://www.insee.fr/fr/metadonnees/definition/c2060.*

La possibilité d'étudier de manière détaillée l'impact de chocs économiques, tels que des variations brutales de prix ou de revenus, est un autre atout majeur de ces données. On peut ainsi estimer, par exemple, dans quelle mesure les ménages réagissent à une hausse de prix en puisant dans leur épargne ou en réduisant leur consommation. L'Insee a pu étudier grâce à ces données les effets budgétaires, redistributifs et environnementaux des remises sur le prix des carburants après les hausses liées à la guerre en Ukraine (Adam et al., 2023). Ce potentiel très riche d'analyse tient en grande partie à la nature même des données collectées, qui incluent les soldes de comptes, le détail et les dates des opérations, ainsi que certaines informations sociodémographiques. Les banques proposent de surcroît une typologie des dépenses, grâce à une catégorisation des paiements par carte selon le type d'établissement bénéficiaire, ce qui se révèle précieux pour analyser la composition de la consommation.

Ces nombreux avantages n'éludent pas les limites des données, à commencer par un problème de représentativité :

- Les personnes non bancarisées ne sont pas prises en compte. Par ailleurs, l'observation des seuls clients d'une banque spécifique, même si cette clientèle couvre un large spectre de la population, est sensible aux effets de spécialisation de chaque établissement.

- La vision de la situation financière est partielle. En effet, les clients sont parfois multibancarisés et peuvent détenir d'autres revenus ou avoir recours à d'autres canaux de dépenses qui échappent à l'analyse. Les dimensions comme le patrimoine immobilier ne sont pas non plus accessibles.

- À ces biais s'ajoute la difficulté d'établir des correspondances exactes avec les concepts usuels de la statistique publique. La notion de ménage doit ainsi être approchée par celle de groupe familial, aboutissant à une mesure incomplète des revenus et des dépenses, si certains membres du ménage ont un compte dans une autre banque. De plus, certaines transactions ne traduisent pas de véritables opérations de consommation ou de perception de revenus, par exemple des transferts entre comptes d'un même individu entre des banques différentes.

- Ces indicateurs peuvent se révéler volatils et induire des conclusions hâtives démenties ensuite lorsque l'environnement économique est moins turbulent.

Plusieurs défis se posent pour maximiser le potentiel de ces données et améliorer la qualité des analyses. Le premier consiste à accroître la représentativité en développant de nouveaux partenariats avec davantage d'établissements afin, d'une part, de potentiellement mieux cerner les personnes multibancarisées et, d'autre part, d'englober un éventail plus large de ménages. Le second concerne la profondeur historique, puisqu'un historique plus ancien permet de mieux appréhender l'évolution des comportements financiers sur le long terme. Une fois ces conditions réunies, la désaisonnalisation[12] prendra tout son sens, car elle contribuera à atténuer une partie du bruit[13] inhérent à la haute fréquence d'observation et facilitera la détection de tendances conjoncturelles.

12 Désaisonnaliser consiste à appliquer un traitement statistique pour éliminer les effets dus aux phénomènes saisonniers.
13 Fluctuations non directement liées au phénomène que l'on cherche à analyser.

Les efforts d'amélioration de la catégorisation des flux financiers constitueront également un atout précieux pour rapprocher progressivement les transactions des définitions usuelles de la statistique publique. Enfin, la préservation de la confidentialité des données demeure centrale, afin de garantir l'anonymat des clients et de protéger leur vie privée. L'objectif à long terme est de parvenir à combiner ces données avec d'autres sources pour disposer d'une vision plus exhaustive du comportement économique des ménages, tout en maintenant des standards élevés de fiabilité et de confidentialité.

Les données de location immobilière de courte durée

Début 2020, Eurostat a conclu un accord d'échanges de données avec quatre plateformes de location de courte durée : *Airbnb, Booking, Expedia Group* et *TripAdvisor*. La collecte de données auprès des plateformes permet d'améliorer la qualité des statistiques européennes sur le tourisme. En effet, s'agissant du marché de l'hébergement, le segment des locations de vacances de courte durée n'est traditionnellement couvert que partiellement. Les acteurs de ce marché de location sont nombreux, ce qui rend la collecte de données plus difficile. Une partie importante des locations de vacances de la nomenclature d'activité européenne « vacances et autres hébergements de court séjour » (code NACE[14] 55.2) n'est ainsi pas représentée.

> ***Les données de location immobilière de courte durée viennent combler les lacunes des enquêtes traditionnelles sur le tourisme, qui ne couvrent pas systématiquement ce segment en plein essor.***

Les données mises à disposition par ces plateformes couvrent les hébergements de courte durée (à l'exclusion des hôtels et campings) réservés par leur intermédiaire dans l'Union européenne (UE) et dans les pays de l'Association européenne de libre-échange (AELE[15]). Les plateformes fournissent trimestriellement à Eurostat des données comme les nombres de nuitées réservées et de voyageurs. Les statistiques obtenues ont un caractère expérimental, comprenant notamment des risques de double compte.

Ces données permettent d'éclairer l'impact économique et social des plateformes de location immobilière de courte durée dans différents pays européens (voir par exemple Ulrich (2021) pour la France). Elles viennent combler les lacunes des enquêtes traditionnelles sur le tourisme, qui ne couvrent pas systématiquement ce segment en plein essor. Leur utilisation est également envisagée pour mieux prendre en compte l'hébergement de courte durée via les plateformes dans le futur système de comptabilité national (Askenazy et Bourgeois, 2025). Ces travaux illustrent en même temps l'importance d'institutions telles qu'Eurostat dans la négociation d'accords avec des acteurs de l'économie numérique qui sont simultanément présents sur différents marchés nationaux. L'échelon européen est, dans ce contexte, mieux adapté pour négocier avec ces acteurs.

14 NACE : nomenclature statistique des activités économiques dans la Communauté européenne : *https://www.insee.fr/fr/metadonnees/definition/c2073* et *https://ec.europa.eu/eurostat/fr/web/nace*.
15 L'AELE compte actuellement quatre pays membres: l'Islande, le Liechtenstein, la Norvège et la Suisse.

▶ Évolutions institutionnelles et juridiques

Avec la publication de la **stratégie européenne pour les données**[16] par la Commission européenne en 2020, l'échelon européen s'est emparé de la question de l'accès aux données détenues par le secteur privé à des fins d'intérêt général. Dans ce contexte, les institutions européennes ont également finalisé, fin 2024, **une révision du règlement n° 223/2009 relatif aux statistiques européennes**[17], via l'adoption d'un règlement rectificatif : le règlement n° 2024/3018 modifiant le règlement n° 223/2009 relatif aux statistiques européennes[18]. Les nouvelles dispositions introduites par cette révision traitent notamment de l'accès aux sources de données privées par le système statistique européen. Ces différents textes précisent par ailleurs ce qu'ils entendent sous le terme de données (***encadré 2***).[19]

▶ Encadré 2. Les définitions des « données » par le droit européen

Dans le cadre de la stratégie européenne pour les données, l'article 2 du *Data Governance Act** définit ce que le règlement entend comme étant une « donnée ». Il s'agit de « toute représentation numérique d'actes, de faits ou d'informations et toute compilation de ces actes, faits ou informations, notamment sous la forme d'enregistrements sonores, visuels ou audiovisuels ». Pour la première fois, un texte européen donne une définition juridique à la notion de « données ».

Pour le statisticien public, cette définition ne va pas nécessairement de soi, les données pouvant exister même si elles n'ont pas de représentation numérique (cas par exemple de tous les anciens registres administratifs sur support papier). Il est d'ailleurs amusant de constater que deux ans et demi plus tard, le règlement n° 2024/3018 introduit dans le règlement n° 223/2009 relatif aux statistiques européennes une définition différente de la donnée : « toute représentation numérique ou non d'actes, de faits ou d'informations et toute compilation de tels actes, faits ou informations sur les unités observées ».

Le droit européen qui, jusqu'à récemment, ne définissait pas ce qu'était une donnée repose désormais sur deux définitions différentes de cette notion. Bien évidemment, la définition applicable à la statistique européenne est celle du règlement n° 223/2009.

* *Voir les références juridiques en fin d'article.*

La stratégie européenne pour les données

La stratégie européenne pour les données comporte deux textes majeurs : le *Data Governance Act* (règlement sur la gouvernance des données)[20] et le *Data Act* (règlement sur les données)[21]. Certains points intéressent en particulier la statistique publique.

Le *Data Governance Act* définit le cadre juridique des **espaces européens de données** (*data spaces*). La Commission européenne soutient leur développement pour faciliter la mise en commun et le partage des données dans des secteurs clés, notamment dans divers champs sectoriels de l'économie (agriculture, santé, énergie, transports, etc.). Si ces espaces se développent et se pérennisent, ils peuvent offrir des perspectives

16 *https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_fr*.
17 Voir les références juridiques en fin d'article.
18 Voir les références juridiques en fin d'article.
19 Au niveau de la France, par ailleurs, un rapport d'information de l'Assemblée nationale propose des évolutions du cadre juridique et des pratiques entourant l'accès aux sources de données privées afin de faciliter leur utilisation par l'Insee (Sala, 2023).
20 Voir les références juridiques en fin d'article.
21 Voir les références juridiques en fin d'article.

intéressantes pour la statistique européenne. D'ores et déjà a été créé en mars 2025 l'espace européen des données de santé[22]. Datant d'avant le *Data Governance Act*, mais dans le même esprit, l'espace de données Copernicus[23] offre un accès aux images satellitaires ainsi qu'un écosystème d'outils et de données dérivées qui en facilitent l'usage[24].

Le *Data Act* porte principalement sur les données générées par l'Internet des objets[25] et sur la concurrence sur le marché du *cloud* (informatique en nuage). En outre, il définit un cadre juridique pour le partage de données des entreprises vers les pouvoirs publics (***encadré 3***). Ces règles concernent l'ensemble des pouvoirs publics, dont bien évidemment la statistique publique.

▶ Encadré 3. Les nouveaux droits des pouvoirs publics introduits par le *Data Act* pour accéder à des données détenues par des entreprises en cas de besoin exceptionnel

Le *Data Act* donne un cadre juridique au partage de données des entreprises vers les pouvoirs publics, communément appelé « B2G » pour *business to government*. Concrètement, les dispositions du chapitre V donnent plus de droits aux pouvoirs publics pour demander des données au secteur privé, mais seulement en cas de nécessité et pour un besoin exceptionnel, ce qui en limite la durée et la portée.

Si le besoin exceptionnel correspond à une situation d'urgence, les données sont mises à disposition gratuitement, et les pouvoirs publics accordent « une reconnaissance publique » aux entreprises qui le demandent.

Hors situation d'urgence, seul l'accès à des données à caractère non personnel est possible pour répondre à un besoin exceptionnel, et uniquement en tout dernier recours : après avoir épuisé tous les autres moyens dont disposent les pouvoirs publics, y compris l'adoption de nouvelles mesures législatives ou l'achat de ces données sur le marché. Lorsque ces conditions sont réunies, le détenteur de données a le droit de demander une compensation visant à couvrir les coûts techniques et organisationnels pour répondre à la demande de transmission de données, ainsi qu'une marge raisonnable.

La révision du règlement n° 223/2009 relatif aux statistiques européennes

La révision du règlement n° 223/2009 relatif aux statistiques européennes introduit d'abord de nouvelles définitions. Parmi celles-ci, un « détenteur de données » est « une personne physique ou morale ou toute autre entité qui a le droit [...] ou la capacité, de gérer et de mettre à disposition des données obtenues dans le cadre de son activité ».

Surtout, **cette révision dote Eurostat et les instituts nationaux de statistique européens de prérogatives nouvelles pour demander l'accès à des sources de données privées à des fins statistiques, en plus de celles prévues par le *Data Act* en cas de besoin exceptionnel**. Le texte s'applique à des données dont il convient de montrer qu'elles sont strictement « nécessaires pour le développement, la production

22 *https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_fr*.
23 *https://www.copernicus.eu/fr/acces-aux-donnees*.
24 De nombreux instituts nationaux de statistique, dont l'Insee, explorent l'usage de l'imagerie spatiale et de la photographie aérienne pour la statistique publique. La plupart de ces données étant publiques, elles ne sont pas évoquées dans cet article.
25 Systèmes interconnectés permettant de superviser et de contrôler des objets à distance grâce à des échanges de données par Internet (Boudrot, 2021).

et la diffusion de statistiques européennes et que celles-ci ne peuvent être obtenues autrement ou que leur réutilisation entraînera une réduction considérable de la charge de réponse pesant sur les détenteurs de données et d'autres entreprises ». Il est important de noter que ces finalités statistiques comprennent les activités scientifiques et de recherche des instituts de statistique ainsi que l'établissement de bases de sondage.

S'agissant d'une potentielle compensation financière, le nouvel article 17 ter du règlement dispose que la mise à disposition des données et métadonnées est gratuite, mais aussi que « [l]orsque les données demandées [...] nécessitent un service de traitement spécifique, les États membres peuvent accorder une compensation au détenteur de données privé pour ce service, sauf lorsque le droit national proscrit l'indemnisation des détenteurs de données ». Par conséquent, la question de la gratuité de l'accès aux sources de données privées est renvoyée au droit de chaque État.

> ***Le nouveau cadre européen invite à établir des relations partenariales avec les détenteurs de données privés.***

Le nouveau cadre défini par cette révision invite aussi à établir des relations partenariales avec les détenteurs de données privés. En effet, après une demande de mise à disposition de données par Eurostat ou un institut national de statistique, un dialogue doit s'engager entre le détenteur de données privé et l'administration ayant formulé la demande « afin de discuter et de convenir des mesures requises pour la mise à disposition des données [...] en vue de conclure un accord ». La demande d'accès ne suffit donc pas, il est nécessaire d'engager une discussion pour trouver un accord. Si aucun accord n'est conclu dans un délai de trois mois, une seconde demande peut être formulée. Elle deviendra alors opposable en droit, c'est-à-dire contraignante pour les détenteurs de données. L'accent mis sur le dialogue entre les acteurs du système statistique européen et les détenteurs de données privés est ainsi proche du mécanisme de concertation qu'a prévu le législateur aux travers des dispositions de l'article 3 bis de la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques[26].

Les nouvelles dispositions du règlement prévoient également qu'Eurostat puisse mettre en place une infrastructure sécurisée pour faciliter le partage de ces données au sein du système statistique européen. Ces dispositions sont en lien avec les investissements qu'Eurostat mène sur les infrastructures de calcul multipartite sécurisé (Ricciato, 2024) au travers notamment du projet JOCONDE[27].

▶ Mettre en place un cadre partenarial soutenable

L'orientation prise par le système statistique européen vise résolument à la mise en place de partenariats avec les détenteurs de données privés. Cette orientation était présente dès 2022 dans les conclusions du groupe européen de haut niveau sur l'utilisation des nouvelles sources de données pour la statistique publique (Eurostat, 2022).

26 Voir les références juridiques en fin d'article.
27 *Joint On-demand COmputation with No Data Exchange* : *https://cros.ec.europa.eu/joconde*.

Cette voie correspond, dans les faits, à celle déjà entreprise dans différents pays européens ainsi qu'en France depuis les années 2010. La principale raison d'être de ces partenariats tient à la nature des données. Aucune source de données privée n'a été conçue à des fins statistiques. Ces sources ont toutes été conçues pour d'autres finalités : gérer un réseau de télécommunications, gérer un système de paiement par carte, offrir un service d'intermédiation et de mise en relation sur internet, gérer des comptes bancaires, etc.

L'investissement nécessaire pour construire des indicateurs qui répondent aux exigences de la statistique publique est conséquent et nécessite un réel travail partenarial. En effet, les statisticiens publics ne peuvent comprendre seuls la complexité que représente la gestion d'un réseau de télécommunications ou d'un système de paiement. Réciproquement, les détenteurs de données privés ne connaissent pas la statistique publique. Les données qu'utilise le service statistique public doivent notamment être alignées sur des référentiels, tels que les nomenclatures statistiques. À cet égard, les données diffusées en open data par la statistique publique participent à l'appropriation de ces référentiels par le secteur privé. Le répertoire Sirene[28] est ainsi très largement exploité par le secteur privé, diffusant par là-même l'utilisation de la nomenclature d'activités française[29] (NAF) au sein de cette sphère.

Par ailleurs, les obligations et exigences de protection de la vie privée auxquelles doivent répondre les détenteurs de données privés représentent autant de freins à la mobilité des données dont ils sont responsables (Desrochers, 2024). Suivant les circonstances, il peut être préférable qu'ils ne transfèrent pas leurs données brutes à l'extérieur de leur système d'information. Dès lors, il peut devenir nécessaire de concevoir et mettre en place des systèmes et des méthodologies qui, à la fois, répondent aux besoins de la statistique publique et prennent en compte ces contraintes. Dans certains cas, un co-investissement de la statistique publique et de ces opérateurs sera la meilleure voie pour trouver des solutions qui répondent aux besoins et contraintes de l'ensemble des partenaires. Dans cette optique, il pourrait être envisagé de recourir à des technologies améliorant la confidentialité ou « PETs » (*Privacy Enhancing Technologies*), telles que la confidentialité différentielle (*differential privacy*) (Tassi, 2019) ou le calcul multipartite sécurisé (Ricciato, 2024). Cependant, ces dernières restent encore particulièrement complexes et coûteuses à mettre en œuvre.

> ***Les accords passés entre le système statistique public et des détenteurs de données privés restent néanmoins fragiles dans le temps.***

Les accords passés entre le système statistique public et des détenteurs de données privés restent néanmoins fragiles dans le temps. La concurrence existant dans le secteur privé peut conduire à un changement de contrôle, voire à la disparition de certains d'entre eux. Les accords jusqu'ici trouvés avec des acteurs privés ont souvent reposé sur une prise de conscience de leur responsabilité sociétale. En ce sens, l'épisode de la crise sanitaire de 2020 a été un catalyseur très positif. Il est néanmoins impossible de présager de l'avenir, et de savoir si

28 *https://www.insee.fr/fr/metadonnees/source/serie/s1020.*
29 *https://www.insee.fr/fr/information/2406147.*

les partenariats avec les instituts de statistique continueront de trouver une place dans leurs stratégies. Les nouvelles prérogatives du système statistique européen – en ce qu'elles permettent de rendre obligatoire la transmission de données qu'ils détiennent – atténuent ces risques.

▶ Les principaux enjeux méthodologiques associés aux sources de données privées

Les sources de données privées proviennent d'acteurs agissant sur un marché concurrentiel. Ces détenteurs de données peuvent donc être nombreux, se différencier sur leurs offres et, partant, leurs clientèles. Dès lors, les données d'un seul acteur ne couvrent pas l'ensemble du champ et peuvent souffrir d'un problème de **représentativité**. Plusieurs stratégies sont alors possibles. On peut chercher à accroître le champ – en augmentant le nombre d'acteurs auprès desquels sont collectées les données – afin qu'il soit le plus large possible (Coudin et al., 2021). Dans certains cas, on peut aussi échantillonner les données ou les caler, si les sources de données comprennent des informations auxiliaires le permettant (Bonnet et Loisel, 2024).

Les sources de données privées souffrent très fréquemment de **différences de concept** au regard de ce que le statisticien cherche à mesurer. Par exemple, en ce qui concerne les données bancaires, les données d'une seule banque ne permettent pas d'avoir une mesure exhaustive de la situation financière des individus du fait du cas des personnes multibancarisées. De même, les paiements par carte bancaire ne mesurent pas les dépenses en espèces ou par chèque.

Le fait que les consommateurs aient recours aux services de plusieurs acteurs concurrents peut conduire à des **comptages multiples**. C'est le cas par exemple, de la téléphonie mobile pour les personnes disposant d'une ligne personnelle et d'une ligne professionnelle, ou bien des clients d'opérateurs étrangers en itinérance qui peuvent être détectés successivement chez différents opérateurs.

Dans certains cas, le **passage à l'unité statistique utilisée par le statisticien** est source de complexité. En effet, il n'y a pas toujours identité entre le souscripteur d'un service et ses utilisateurs : abonnement de téléphonie mobile au nom d'un parent utilisé par un enfant, carte de paiement utilisée par un autre membre de la famille, compte bancaire conjoint, etc.

Afin de corriger la plupart de ces problèmes, il est souvent procédé à des enquêtes pour en évaluer leurs ampleurs. Cela conduit à un paradoxe, car ces nouvelles sources de données ont souvent été présentées comme pouvant réduire le recours à des enquêtes.

▶ Quel avenir, alors, pour ces nouvelles sources ?

La documentation accumulée sur ces enjeux méthodologiques est le résultat des partenariats mis en œuvre dans le service statistique public avec des détenteurs de données privés. La connaissance de ces sources, de leur potentiel, mais aussi de leurs limites est précisément ce qui peut permettre, à terme, leur intégration aux côtés des autres sources mobilisées par le service statistique public.

Ces données privées offrent d'ores et déjà des connaissances inaccessibles à l'aide d'enquêtes. Elles ne permettent pas pour autant de s'y substituer, chaque source ayant ses forces et faiblesses.

Ces données privées offrent d'ores et déjà des connaissances inaccessibles à l'aide d'enquêtes. Elles ne permettent pas pour autant de s'y substituer, chaque source ayant ses forces et faiblesses. Les enquêtes restent ainsi le seul moyen de couvrir un champ de façon exhaustive. Par ailleurs, même si les prérogatives des instituts nationaux de statistique ont été renforcées par la révision du règlement n° 223/2009, les accès aux sources de données privées reposent sur des accords qui n'offrent pas la même garantie de pérennité. Aujourd'hui, elles apparaissent donc bien comme des sources complémentaires, qui peuvent trouver leur place aux côtés des enquêtes et des sources administratives déjà exploitées par le service statistique public, et permettre ainsi d'enrichir la production statistique.

Comme Harford l'avait expliqué dès 2014, ces nouvelles traces numériques des entreprises privées ne sont pas la panacée, ne peuvent se substituer aux enquêtes déjà existantes, et ne peuvent faire l'économie d'une analyse méthodologique approfondie (Harford, 2014). Mais elles permettent d'offrir de nouveaux éclairages pour enrichir la connaissance de la société et de l'économie française.

▶ Fondements juridiques

- *Data Governance Act* : règlement (UE) nº 2022/868 du Parlement européen et du Conseil du 30 mai 2022 portant sur la gouvernance européenne des données et modifiant le règlement (UE) nº 2018/1724. In : *site de l'Union européenne*. [en ligne]. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://eur-lex.europa.eu/eli/reg/2022/868/oj*.

- *Data Act* : règlement (UE) nº 2023/2854 du Parlement européen et du Conseil du 13 décembre 2023 concernant des règles harmonisées portant sur l'équité de l'accès aux données et de l'utilisation des données et modifiant le règlement (UE) nº 2017/2394 et la directive (UE) nº 2020/1828. In : *site de l'Union européenne*. [en ligne]. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://eur-lex.europa.eu/eli/reg/2023/2854/oj.*

- Règlement (UE) nº 2024/3018 du Parlement européen et du Conseil du 27 novembre 2024 modifiant le règlement (CE) nº 223/2009 relatif aux statistiques européennes. In : *site de l'Union européenne*. [en ligne]. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://eur-lex.europa.eu/eli/reg/2024/3018/oj*.

- Loi nº 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques. In : *site de Légifrance*. Mise à jour le 25 mars 2019. [en ligne]. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573*.

▶ Bibliographie

- ADAM, Marine, BONNET, Odran, FIZE, Étienne, RAULT, Marion, LOISEL, Tristan, WILNER, 2023. L'ajustement de court terme de la consommation de carburant à des changements de prix – Des estimations menées à partir de données à haute fréquence. In : *Documents de travail*. [en ligne]. 7 juillet 2023. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/7644834*.

- ASKENAZY, Philippe et BOURGEOIS, Alexandre, 2025. Vers une meilleure prise en compte de l'hébergement via des plates-formes en ligne au sein des comptes nationaux. In : *Documents de travail*. [en ligne]. 4 mars 2025. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/8376381*.

- BLANCHET, Didier et GIVORD, Pauline, 2017. Données massives, statistique publique et mesure de l'économie. In : *L'économie française, coll. « Insee Références »*. [en ligne]. 11 juillet 2017. Pp. 59-77. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/2894010?sommaire=2894036*.

- BONNET, Odran et LOISEL, Tristan, 2024. L'économie racontée par les données bancaires – Ce que nos relevés de comptes disent de nous. In : *Courrier des statistiques*. [en ligne]. 16 décembre 2024. Insee. N° N12, pp. 115-136. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8264558?sommaire=8264562*.

- BORTOLI, Clément et COMBES, Stéphanie, 2015. Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées. In : *Note de conjoncture*. [en ligne]. 2 avril 2015. Insee. Pp. 43-56. [Consulté le 20 mai 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/1408926?sommaire=1408931*.

- BOUDROT, Nicolas, 2021. Internet des objets, impression 3D, robotique : des technologies davantage utilisées par les grandes sociétés. In : *Insee Première*. [en ligne]. 21 avril 2021. Insee. N° 1854. [Consulté le 12 mai 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/5356737*.

- COMTE, Frédéric, DEGORRE, Arnaud et LESUR, Romain, 2022. Le SSPCloud : une fabrique créative pour accompagner les expérimentations des statisticiens publics. In : *Courrier des statistiques*. [en ligne]. 20 janvier 2022. Insee. N° N7, pp. 68-87. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6035940?sommaire=6035950*.

- COTTON, Franck et HAAG, Olivier, 2023. L'intégration des données administratives dans un processus statistique – Industrialiser une phase essentielle. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp. 104-125. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635829?sommaire=7635842*.

- COUDIN, Élise, POULHES, Mathilde et SUAREZ CASTILLO, Milena, 2021. The French official statistics strategy: Combining signaling data from various mobile network operators for documenting COVID-19 crisis effects on population movements and economic outlook. In : *Data & Policy*. [en ligne]. 24 juin 2021. Vol. 3, p. e10. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://doi.org/10.1017/dap.2021.1*.

- DESROCHERS, Pierre R., 2024. Access to Information and Privacy: Practical Approaches for Public Service Reform. In : *Canadian Public Administration*. [en ligne]. 19 décembre 2024. Volume 67, Issue 4. Pp. 562-572. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://doi.org/10.1111/capa.12582*.

- DONDON, Alexis et LAMARCHE, Pierre, 2023. Quels formats pour quelles données ? In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp. 86-103. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635827?sommaire=7635842*.

- EUROSTAT, 2022. Empowering society by reusing privately held data for official statistics: final report prepared by the high level expert group on facilitating the use of new data sources for official statistics — A European approach — 2022 edition. In : *site de Eurostat*. [en ligne]. Publications Office of the european Union. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/web/products-statistical-reports/-/ks-ft-22-004*.

- EUROSTAT, 2023. Reusing Mobile Network Operator data for Official Statistics: the case for a common methodological framework for the European Statistical System, 2023 edition. In : *site de Eurostat*. [en ligne].[Consulté le 16 mai 2025]. Publications Office of the european Union. Disponible à l'adresse : *https://ec.europa.eu/eurostat/web/products-statistical-reports/w/ks-ft-23-001*.

- GALIANA, Lino, SAKAROVITCH, Benjamin, SÉMÉCURBE, François et SMOREDA, Zbigniew, 2020. Évolution de la ségrégation pendant la journée et frictions spatiales : une analyse à partir de données de téléphonie mobile. In : *Documents de travail*. [en ligne]. 9 novembre 2020. Insee. N° G2020-12. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4925200*.

- GALIANA, Lino et SUAREZ CASTILLO, Milena, 2022. Fuzzy matching on big-data: an illustration with scanner and crowd-sourced nutritional datasets. In : *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*. [en ligne] Pp. 331-337. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://doi.org/10.1145/3524458.3547244*.

- HARFORD, Tim, 2014. Big data: are we making a big mistake? In : *Financial Times*. [en ligne]. 28 mars 2014. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0*.

- INSEE, 2019. L'économie et la société à l'ère du numérique. In : *Insee Références*. [en ligne]. 4 novembre 2019. pp 55-69. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4238635*.

- INSEE, 2020. Éclairage - Disparités territoriales de consommation : que disent les données de transaction par carte bancaire ? In : *Notes et points de conjoncture de l'année 2020*. [en ligne]. 15 décembre 2020. [Consulté le 29 janvier 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4488582?sommaire=4473296&q=cartes+bancaires*.

- LE SAOUT, Ronan, RIEDINGER, Nicolas et MESQUI, Bérengère, 2024. Les statistiques publiques de l'énergie – Enjeux passés, présents et futurs. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2024. Insee. N° N11, pp. 51-71. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8203038*.

- LECLAIR, Marie, 2019. Utiliser les données de caisses pour le calcul de l'indice des prix à la consommation. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 61-75. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254225?sommaire=4254170*.

- RICCIATO, Fabio, 2024. Steps Toward a Shared Infrastructure for Multi-Party Secure Private Computing in Official Statistics. In : *Journal of Official Statistics*. [en ligne]. 15 mars 2024. Volume 40, Issue 1. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://doi.org/10.1177/0282423X241235259*.

- SALA, Michel, 2023. Rapport d'information sur l'accès aux données privées : une nouvelle ressource pour l'Institut national de la statistique et des études économiques ? In : *Rapport d'information de l'Assemblée Nationale*. [en ligne]. N° 1312. 1er juin 2023. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.assemblee-nationale.fr/dyn/16/rapports/cion_fin/l16b1312_rapport-information#_Toc256000048*.

- SAKAROVITCH, Benjamin, DE BELLEFON, Marie-Pierre, GIVORD, Pauline et VANHOOF, Maarten, 2019. Estimer la population résidente à partir de données de téléphonie mobile, une première exploration. In : *Économie et Statistique / Economics and Statistics*. [en ligne]. 11 avril 2019. N° 505-506. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/3706213?sommaire=3706255*.

- SUAREZ CASTILLO, Milena, SÉMÉCURBE, François, ZIEMLICKI, Cezary, TAO, Haixuan Xavier et SEIMANDI, Tom, 2023. Temporally Consistent Present Population from Mobile Network Signaling Data for Official Statistics. In : *Journal of Official Statistics*. [en ligne]. 10 décembre 2023. Vol. 39, N° 4, 2023, pp. 535–570. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://doi.org/10.2478/jos-2023-0025*.

- TASSI, Philippe, 2019. Introduction – Les apports des Big Data. In : *Économie et Statistique / Economics and Statistics*. [en ligne]. 11 avril 2019. N° 505-506. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/3705956?sommaire=3706255*.

- TAVERNIER, Jean-Luc et OURLIAC, Benoît, 2020. Google en sait-il plus que l'Insee sur les Français ? In : *Le blog de l'Insee*. [en ligne]. 18 décembre 2020. [Consulté le 19 mai 2025]. Disponible à l'adresse : *https://blog.insee.fr/google-en-sait-il-plus-que-linsee-sur-les-francais/*.

- TIGANI, Jordan, 2023. Big Data is Dead. In : *MotherDuck Blog*. [en ligne]. 7 février 2023. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://motherduck.com/blog/big-data-is-dead/*.

- ULRICH, Amandine, 2021. Hébergements proposés par des particuliers via des plateformes – En 2019, Paris et Nice dans le top 10 des villes les plus fréquentées de l'Union européenne. In : *Insee Première*. [en ligne]. 26 novembre 2021. Insee. N° 1879. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/5871113*.

• VACHER, Thomas et PRADINES, Nadège, 2017. Cloud computing, big data : de nouvelles opportunités pour les sociétés. In : *Insee Première*. [en ligne]. 30 mars 2017. Insee. N° 1643. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/2672067*.

Les données de téléphonie mobile

Une source de connaissance sur la population et ses déplacements

*Marie-Pierre Joubert**

Les données de téléphonie mobile offrent des perspectives uniques pour la statistique publique, notamment pour étudier la présence de la population sur le territoire. Ainsi, grâce à des partenariats avec des opérateurs (Orange, Bouygues, SFR), l'Insee a pu analyser très rapidement les déplacements de population pendant la crise de la Covid-19, fournissant des informations cruciales pour calibrer les services publics. Ces données, collectées via la signalisation du téléphone sur le réseau, permettent de suivre les mobilités avec une précision temporelle fine. Cependant, elles posent des défis méthodologiques, au sujet en particulier de la couverture géographique et de la représentativité. Les collaborations avec les opérateurs, sous des cadres législatifs stricts aux niveaux national et européen, permettent d'accéder à des informations anonymisées et agrégées. Ces dernières complètent les sources traditionnelles, offrant une vision plus fine des dynamiques de population, importantes pour certaines politiques publiques.

Mobile network data provides exceptional opportunities for official statistics, especially for studying the population's presence in a given area. Thanks to partnerships with operators (Orange, Bouygues, SFR), INSEE was able to quickly analyse population mobility during the Covid-19 crisis, providing essential information for adjusting public services. These data, collected via network signaling, enable mobility to be tracked with a high degree of temporal precision. However, they raise methodological challenges, particularly in terms of geographical scope and representativeness. Collaboration with operators, under strict national and European legislative frameworks, provides access to anonymised and aggregated information. This complements traditional sources, providing a more detailed view of population dynamics, which are important for various public policies.

* À la date de rédaction de l'article, responsable adjointe du SSP Lab, Insee.
marie-pierre.joubert@finances.gouv.fr

Dans quelle mesure la population s'est-elle confinée ailleurs que dans sa résidence principale pendant la crise de la Covid-19, par exemple chez de la famille, des amis ou dans une résidence secondaire ? La réponse à cette question était cruciale en cette période pour calibrer les services publics, notamment sanitaires. Or, les données habituellement utilisées par la statistique publique pour localiser la population (recensement de la population, sources fiscales) renseignent sur l'adresse des résidences principales et secondaires, mais pas sur leur occupation en temps réel. Grâce à un partenariat avec trois opérateurs de téléphonie mobile (Orange, Bouygues, SFR), l'Insee a pu mobiliser les données collectées par ces derniers grâce aux signaux envoyés par les téléphones sur leurs réseaux, pour analyser les déplacements de population pendant la crise sanitaire (Sémécurbe et al., 2020). Il a pu ainsi fournir aux préfectures un précieux éclairage sur le nombre de personnes présentes sur leur territoire.

Cet exemple marquant d'usage des données de téléphonie mobile s'inscrit dans le contexte de multiples collaborations entre l'Insee et les opérateurs : partenariats de recherche, projets européens grâce auxquels les instituts statistiques de plusieurs pays collaborent pour élaborer des méthodologies de combinaison des données, projets en lien avec le milieu académique financés par l'Agence nationale de la recherche, etc. Le présent article décrit les enjeux méthodologiques liés à l'utilisation de ces données, ainsi que les diverses modalités permettant au statisticien public d'y accéder, avec leurs avantages et inconvénients respectifs. Il conclut sur les perspectives offertes, notamment au niveau européen.

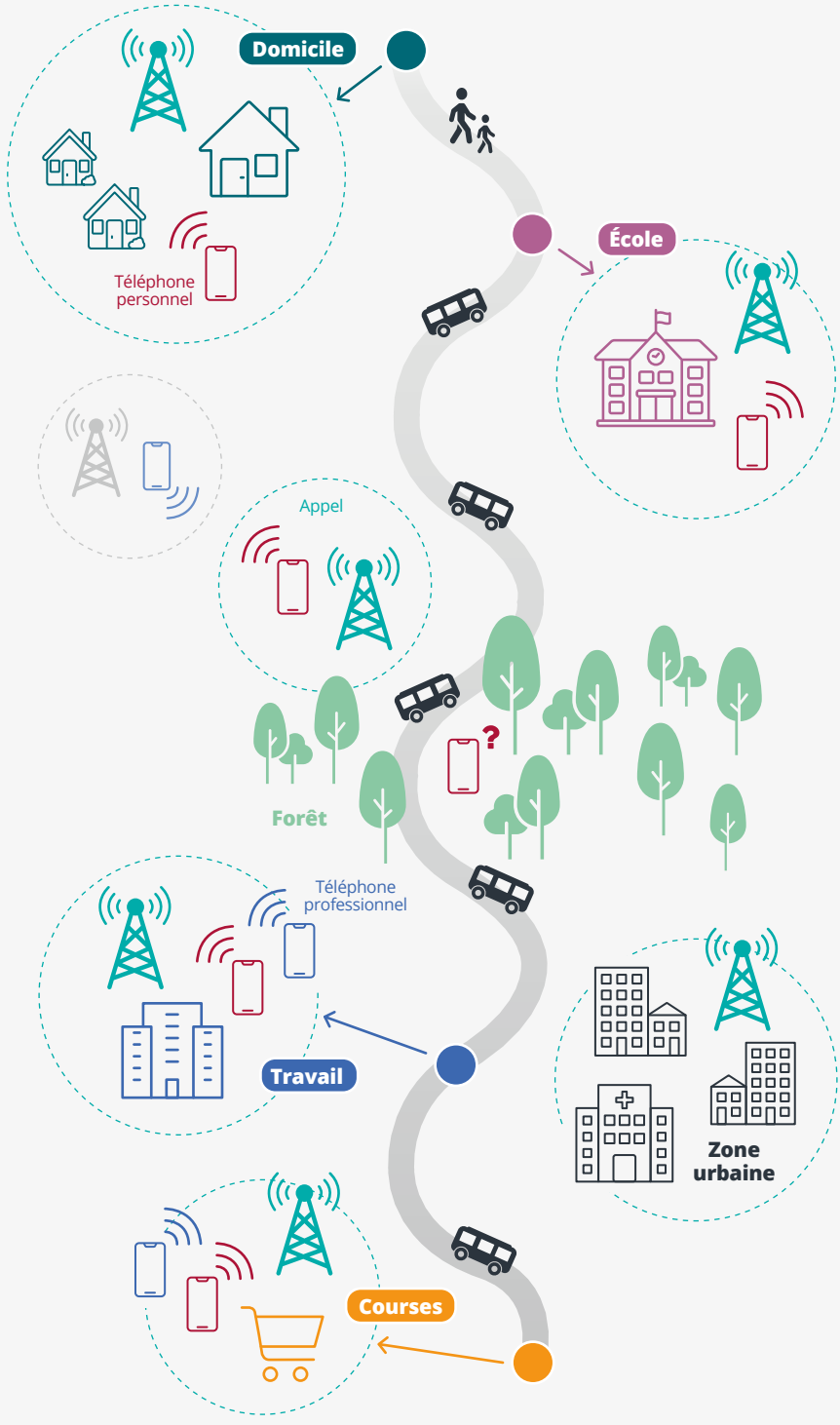
▶ Du téléphone mobile à la base de données : que collectent les opérateurs ?

Ce matin à 7h, madame Zaoui allume son téléphone en se réveillant. À 8h, elle dépose ses enfants à l'école, puis prend le bus vers son travail. En chemin, elle appelle son frère. La communication s'interrompt au moment où elle traverse la forêt, mais reprend dès qu'elle rejoint la zone urbaine. Elle passe ensuite la journée au bureau. Elle profite de la pause méridienne pour aller faire quelques courses au centre commercial le plus proche ; elle emporte alors avec elle son téléphone professionnel, qu'elle a allumé en arrivant au bureau (***figure 1***). Quelles informations sont collectées par l'opérateur téléphonique et consignées dans ses bases de données ?

Les comptes rendus d'appels

Les comptes rendus d'appel de facturation ou *Billing Call Detail Records* (CDR) sont enregistrés à chaque fois qu'un usager utilise le réseau mobile pour envoyer ou recevoir un appel ou un SMS, ou que les applications embarquées dans le mobile transfèrent des données. Autrement dit, il s'agit de tout évènement actif (à la « demande » du mobile) qui laisse une trace de facturation. Ici, un CDR est donc enregistré quand madame Zaoui appelle son frère. Il contient l'heure de l'appel, sa durée et la localisation des antennes les plus proches des téléphones de madame Zaoui et de son frère durant leur communication. Les données sont anonymisées, en conformité avec le règlement général sur la protection des données (RGPD). Par ailleurs, **le contenu de la**

▶ Figure 1 - Suivi des connexions téléphoniques de madame Zaoui



conversation ne fait pas partie des données enregistrées : seules les métadonnées[1] de l'appel sont conservées.

Dans le cadre des CDR, aucune information n'est enregistrée lorsque l'utilisateur se déplace en portant simplement son téléphone sans interagir avec le réseau. L'information est donc parcellaire d'un point de vue temporel. Sakarovitch et al. (2019) soulignent l'impact de cette irrégularité temporelle sur la production de statistiques sur la population.

Les données de signal

Les données de signalisation complètes (données de signal ou *signaling data*) comprennent également des informations liées à l'itinérance du téléphone mobile, qui visent à garantir la qualité de connexion au réseau. En effet, en dehors de tout évènement actif, si le téléphone se déplace significativement, il se signale automatiquement afin d'être joignable sur sa nouvelle position.

Les données de signal contiennent l'identifiant de l'antenne la plus proche du téléphone. Leur grand avantage par rapport aux CDR est leur fréquence temporelle plus élevée. Ainsi, selon Bonnetain et al. (2021), le temps médian entre deux évènements de signalisation est inférieur à 1 minute pour 90 % des utilisateurs, tandis qu'il est de plus de 30 minutes dans les CDR[2].

Contrairement aux CDR, cependant, les données de signal ne contiennent pas d'information sur les contacts entre usagers. Par ailleurs, les personnes comme madame Zaoui qui possèdent deux téléphones, un personnel et un professionnel, risquent d'être comptées deux fois. Ce phénomène est difficile à prendre en compte et fait partie des nombreuses difficultés méthodologiques à résoudre pour produire des statistiques publiques à partir de ces données.

▶ Quelle que soit la maille d'agrégation des données, de nombreuses difficultés méthodologiques à résoudre

Les principales incertitudes inhérentes aux données de téléphonie mobile sont l'incertitude temporelle, l'incertitude sur la couverture de la population et l'incertitude spatiale (Ricciato et al., 2020). L'incertitude temporelle est surtout problématique lors du traitement des CDR. En effet, la fréquence temporelle à laquelle ces dernières fournissent une information dépend de l'usage du réseau téléphonique par l'usager. Dans le cas des données de signal, il est plus facile de compléter les informations manquantes entre deux remontées de données puisque celles-ci sont plus nombreuses. Les deux autres types d'incertitudes concernent à la fois les CDR et les données de signal. D'autres questions méthodologiques s'ajoutent par ailleurs, que doit résoudre le statisticien public pour envisager d'utiliser ces données.

1 Une métadonnée est une donnée qui fournit de l'information sur une autre donnée.
2 Avec le développement des applications mobiles qui échangent des données, la différence entre CDR et données de signal a tendance à s'atténuer. En effet, les CDR intègrent aussi ces échanges de données, dont la fréquence est souvent élevée.

Gérer l'incertitude sur les caractéristiques de la population des abonnés de l'opérateur…

> ***Travailler avec un unique opérateur peut conduire à des imprécisions ou biais dans les décomptes de population.***

La couverture du territoire n'est pas identique suivant les opérateurs (ARCEP, 2024). Par ailleurs, les caractéristiques sociodémographiques de leurs abonnés sont en général différentes. Travailler avec un unique opérateur peut ainsi conduire à des imprécisions ou biais dans les décomptes de population pour certaines zones géographiques où la couverture réseau serait plus faible ou la population moins abonnée à cet opérateur spécifique.

Lorsque le redressement des données n'est pas effectué par l'opérateur, le statisticien public a besoin d'obtenir des informations sociodémographiques sur les clients de façon à pouvoir redresser les données. Or, pour ne pas divulguer le secret des affaires, l'opérateur ne diffuse pas toujours cette information à une échelle géographique fine, ce qui diminue la qualité des redressements effectués.

… et identifier des zones de domicile

> ***Un élément-clé pour le statisticien travaillant sur les données de téléphonie mobile est l'identification de la zone de domicile la plus probable.***

Un élément-clé pour le statisticien travaillant sur les données de téléphonie mobile est l'identification de la zone de domicile la plus probable. Cette information est cruciale pour faire le lien avec les sources de la statistique publique, comme le recensement de la population, et estimer ainsi la représentativité des données, voire recaler ces dernières sur les chiffres du recensement. Le domicile est également un prérequis à de nombreuses analyses, par exemple celle des déplacements domicile-travail. Diverses méthodes d'attribution de domicile existent et les résultats peuvent différer selon les méthodes, ce qui introduit là encore de l'incertitude.

Pour les données issues des CDR, Vanhoof et al. (2018) comparent cinq algorithmes classiques qui identifient respectivement le domicile comme le lieu :

- où la majorité des activités téléphoniques (appels ou SMS, émis ou reçus) ont été effectués ;
- où le nombre maximum de jours distincts avec des activités téléphoniques a été observé ;
- où la plupart des activités téléphoniques ont été enregistrées entre 19h00 et 9h00 ;
- où la plupart des activités téléphoniques ont été enregistrées en considérant un périmètre de 1 000 mètres autour d'une antenne-relais ;
- où la plupart des activités téléphoniques ont été enregistrées entre 19h00 et 9h00 en considérant un périmètre de 1 000 mètres autour d'une antenne-relais.

À un niveau agrégé, quel que soit l'algorithme, la différence de répartition de la population avec celle obtenue à partir du recensement n'est pas négligeable, ce qui illustre les difficultés méthodologiques posées par ces données. Vanhoof et al. (2018) insistent sur l'importance de créer un jeu de données ad hoc permettant de valider les résultats, par exemple en disposant, pour un échantillon d'utilisateurs, à la fois de leur domicile réel et des traces laissées par leurs communications téléphoniques. Connaitre la répartition des parts de marché de l'opérateur est également une information très importante pour la qualité de l'attribution du domicile, de même qu'en savoir plus sur les habitudes d'usage du téléphone.

Pour les données de signal, Suarez Castillo et al. (2023) identifient la zone de domicile grâce à un algorithme qui s'appuie sur la fréquence et la localisation des signaux enregistrés. Ils agrègent ensuite ces données à un niveau spatial fin. Ils les comparent alors à celles de la population résidente estimée à ce même niveau à partir du dispositif Filosofi[3], issu du rapprochement des données fiscales et des données sur les prestations sociales. Cette comparaison leur permet de déduire des poids pour redresser les données de téléphonie mobile. De premiers résultats prometteurs laissent augurer de futures avancées méthodologiques.

Gérer l'incertitude spatiale en estimant la couverture de l'antenne

Du point de vue spatial, disposer uniquement de l'information sur la localisation de l'antenne la plus proche apporte beaucoup d'incertitude, surtout dans les zones rurales peu couvertes en antennes (Sakarovitch et al., 2019). En effet, dans cette situation, on approxime la couverture géographique de l'antenne par un polygone selon la méthode inventée par le mathématicien russe Gueorgui Voronoï[4].

> ***Les opérateurs disposent des connaissances de radio-ingénierie leur permettant d'estimer de façon fiable la couverture de l'antenne. Suivant les modalités de collaboration avec eux, ils peuvent mettre à disposition ces informations.***

Or la couverture réelle est souvent éloignée de cette représentation : elle dépend de la nature de l'antenne, des bâtiments alentours et même de la météo.

Des méthodes probabilistes permettent d'améliorer un peu la précision de l'estimation (Salgado et al, 2021 ; Ricciato et Coluccia, 2021 ; Gootzen et Tennekes, 2022). Bonnetain et al. (2021) ont par ailleurs développé un algorithme permettant d'augmenter significativement la précision spatiale des déplacements reconstruits par la téléphonie mobile dans un environnement urbain.

3 *https://www.insee.fr/fr/metadonnees/source/serie/s1172.*
4 Voronoï a élaboré un algorithme mathématique qui permet, partant d'un ensemble discret de points, de partitionner l'espace en polygones autour de ces points avec la propriété suivante : pour un point p de l'ensemble discret de points P, tous les points contenus dans le polygone associé à p sont plus proches de p que d'aucun autre point de l'ensemble P.

Toutefois, les opérateurs disposent des connaissances de radio-ingénierie leur permettant d'estimer de façon fiable la couverture de l'antenne. Suivant les modalités de collaboration avec eux, ils peuvent mettre à disposition ces informations. Une incertitude demeure dans certains cas où la couverture d'une antenne intersecte celle de ses voisines, mais cela reste plus précis que l'estimation à partir des polygones de Voronoï.

Les questions méthodologiques selon le niveau d'agrégation des données

Sous réserve du respect de conditions strictes garantissant la non-divulgation de la vie personnelle des individus, ainsi que la sécurité des données, **certains partenariats permettent aux instituts de statistique publique d'accéder à des données individuelles pseudonymisées**. Ces données contiennent une ligne par identifiant de téléphone et minute de connexion au réseau, avec l'information sur la localisation de l'antenne la plus proche du téléphone. S'il s'agit de CDRs, les informations sur l'identifiant du téléphone contacté et la localisation de l'antenne la plus proche de ce téléphone sont également disponibles. Ces données sont de **taille massive**. Les traiter nécessite donc une infrastructure de stockage spécifique et des méthodes de science des données (ou *data science*) adaptées. La plupart du temps, les statisticiens se déplacent dans les locaux de l'opérateur pour travailler directement sur les serveurs sécurisés.

Les données individuelles ont l'avantage d'être exhaustives et de permettre à l'analyste de maîtriser de bout en bout les traitements effectués. La méthodologie est ainsi intégralement connue par l'analyste et des statistiques descriptives à façon peuvent être calculées. Ceci ne lève pas pour autant toutes les difficultés méthodologiques précitées, comme celles liées à l'évaluation des parts de marché de l'opérateur ou à la couverture spatiale de l'antenne.

Dans d'autres types de partenariats, l'opérateur fournit directement des données agrégées. L'opérateur a ainsi procédé lui-même au redressement des données pour les rendre représentatives de l'ensemble de la population résidant en France, ainsi qu'à leur projection géographique sur des zonages administratifs (commune, Iris[5]). Ces données sont en général d'une taille raisonnable, les traitements en sont donc grandement facilités. Toutefois, le plus souvent, les opérateurs ne divulguent pas l'intégralité des traitements méthodologiques effectués et tronquent les données, par exemple en ne diffusant pas d'information sur les déplacements de commune à commune représentant moins de 20 individus. Ceci restreint les usages possibles et rend la statistique publique dépendante d'un travail en amont qu'elle ne maîtrise pas.

La plupart du temps, les informations mises à disposition consistent en des décomptes de populations ayant été présentes dans une zone géographique donnée durant une période de temps donnée, par exemple une demi-heure. Il s'agit également de matrices origine-destination indiquant le nombre de personnes s'étant rendues d'une commune A à une commune B pendant un intervalle de temps donné. Ces informations permettent d'étudier la population présente et les déplacements de population. Elles nécessitent toutefois un cadrage méthodologique supplémentaire. Par exemple, faut-il définir une durée d'arrêt

5 L'Iris constitue la brique de base en matière de diffusion de données infracommunales. *https://www.insee.fr/fr/metadonnees/definition/c1523*.

> ***La plupart du temps, les informations mises à disposition consistent en des décomptes de populations ayant été présentes dans une zone géographique donnée durant une période de temps donnée, par exemple une demi-heure.***

minimal dans une zone pour considérer qu'une personne y a été présente ? Ou encore, une personne doit-elle être comptée deux fois dans une zone si elle l'a traversée plusieurs fois ?

Les données sont souvent segmentées suivant la « zone de nuitée » qui peut être assimilée au domicile de l'individu et est directement calculée par l'opérateur. Aude et al. (2024) utilisent ces données agrégées en complément des données traditionnelles de l'Insee pour décrire le fonctionnement des territoires. Les quartiers de Lyon y sont d'abord caractérisés par une typologie basée sur leur population et leur parc de logements, obtenues grâce aux sources de données traditionnelles de la statistique publique. Puis leur fréquentation en journée est analysée grâce aux données de téléphonie mobile.

Que ces données soient individuelles ou agrégées, y accéder nécessite d'établir un partenariat entre un ou plusieurs opérateur(s) de téléphonie mobile, un ou plusieurs institut(s) de statistique publique et parfois d'autres acteurs, par exemple issus du milieu académique.

▶ Des modalités d'accès qui concilient respect de la vie privée, qualité des informations et enjeux commerciaux des opérateurs

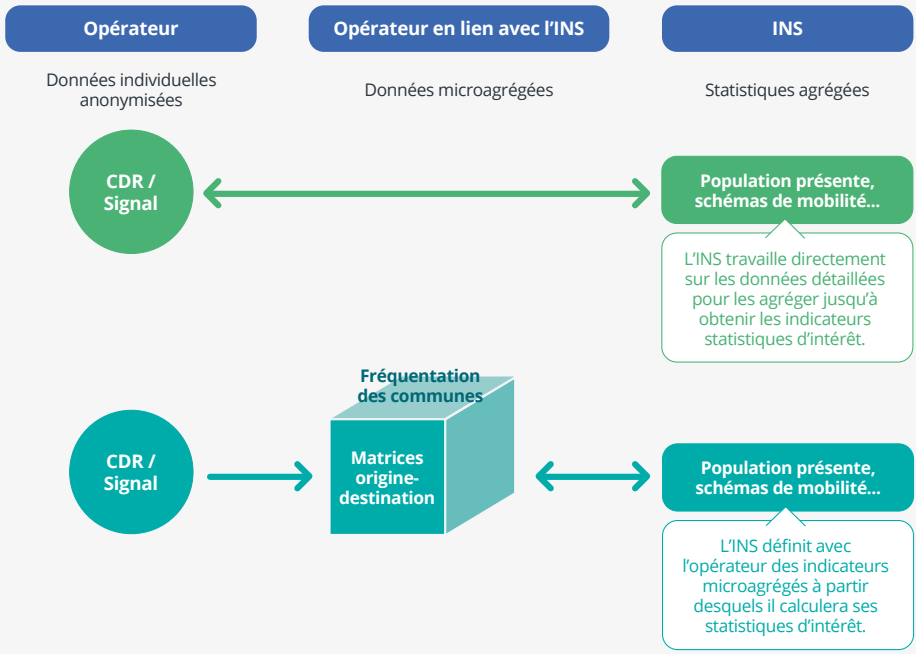
L'Insee a expérimenté plusieurs types de partenariats…

À ce jour, **deux principaux types de partenariat ont été expérimentés par l'Insee** : l'un permet d'accéder aux données détaillées, l'autre aux données agrégées (***figure 2***).

L'accès aux données détaillées a été obtenu pour la période 2016-2022 dans le cadre d'une convention tripartite entre Orange-Labs (le laboratoire de recherche en sociologie d'Orange), l'Insee et Eurostat[6]. Chaque partie permettait aux autres d'accéder à ses bases détaillées dans ses propres locaux. L'opérateur Orange apportait son expertise méthodologique et analytique de l'utilisation de la téléphonie mobile pour l'analyse sociale et économique et fournissait toutes les informations nécessaires pour utiliser les données. L'Insee assurait la disponibilité de sources susceptibles d'être utiles pour améliorer et évaluer la qualité des statistiques dérivées de la téléphonie mobile. Des accords de confidentialité s'assuraient de la non-diffusion d'informations sensibles issues des bases de l'un ou l'autre des partenaires. Plusieurs publications de recherche ont résulté de ce partenariat ; par exemple, l'une d'entre elles confronte le zonage en aires urbaines aux données de téléphonie mobile (Combes et al., 2017).

6 Eurostat est l'Office statistique de l'Union européenne.

Figure 2 - Modalités d'accès des instituts nationaux de statistique aux données de téléphonie mobile



CDR : Comptes rendus d'appel ou *Call Detail Records*.
INS : Institut national de statistique.

> ***Dans le contexte exceptionnel de la crise de la Covid-19, trois des quatre opérateurs majeurs ont répondu favorablement aux sollicitations de l'Insee et engagé des collaborations philanthropiques et limitées dans le temps.***

Dans le contexte exceptionnel de la crise de la Covid-19, trois des quatre opérateurs majeurs ont répondu favorablement aux sollicitations de l'Insee et engagé des collaborations philanthropiques et limitées dans le temps. Des accords de confidentialité ou bons de commande ont été établis pour encadrer la livraison et l'utilisation des données agrégées. Disposer des données de trois opérateurs a permis d'améliorer la qualité des statistiques produites, qui ont été diffusées dans deux communiqués de presse et une publication *Insee Analyses* (Galiana et al., 2020). Toutefois, de façon à ne pas révéler leurs parts de marché respectives à une échelle géographique fine, les opérateurs ont fourni des données déjà ajustées pour les rendre représentatives de la population. Cela a rendu difficile le retraitement par l'Insee, de même que la documentation des limites méthodologiques.

Un troisième type de partenariat est actuellement en cours. Il s'agit d'un partenariat de recherche, financé par l'Agence nationale de la recherche, entre l'université Gustave Eiffel, Orange et l'Insee. Le projet a pour objectif d'utiliser des données de téléphonie mobile,

combinées à d'autres sources de données, pour estimer en continu des indicateurs de présence et de mobilité des personnes, sur la zone du Grand Lyon. Le service commercial d'Orange met à disposition des partenaires des données agrégées qui correspondent au besoin du projet (présence et matrices origine-destination). Les équipes de recherche et développement d'Orange participent aux réunions et apportent l'ensemble des éléments méthodologiques nécessaires à la bonne compréhension des données. De plus, un post-doctorant travaille sur les bases détaillées dans les locaux d'Orange pour concevoir de nouveaux indicateurs expérimentaux. Enfin, une enquête ad hoc est menée sur un échantillon de volontaires, pour confronter leurs trajectoires réelles et les traces repérées par le réseau, ce qui est précieux pour mieux comprendre les données. Ce partenariat, bien que donnant principalement accès à des données agrégées, permet donc de continuer à améliorer la compréhension des données et à illustrer leur intérêt pour la statistique publique. Les récentes avancées législatives au niveau européen ouvrent toutefois la voie à un accès plus généralisé aux données détaillées.

… qui s'inscrivent dans le contexte juridique français et européen…

De manière générale, on entend par **base de données privée** toute base de données collectée ou produite par des organismes de droit privé dans le cadre de leurs activités. Ces données peuvent concerner des tiers, notamment des personnes physiques, et donc être à ce titre protégées par le RGPD, ce qui est le cas pour les données de téléphonie mobile. Elles peuvent être l'objet, de la part de leur détenteur, d'une valorisation à des fins commerciales ou d'une publication. Elles peuvent contenir des informations dont le secret est protégé par la loi et relever pour ce motif des obligations dues au secret professionnel. La circulation de ces données est régie par plusieurs lois imbriquées aux niveaux français et européen, qui protègent donc d'une part la propriété de la base de données (droit d'auteur, de propriété), d'autre part la diffusion du contenu de ces bases, notamment ce qui est lié au caractère de leur contenu (comme les données à caractère personnel).

En France, **la loi du 7 octobre 2016 pour une République numérique**[7] favorise l'ouverture de l'accès aux données publiques (open data) et encourage la circulation des données, la protection des individus dans la société du numérique et l'accès au numérique pour tous. Cette loi introduit un article 3 bis dans la loi du 7 juin 1951[8], qui oblige les personnes morales de droit privé à « transmettre par voie électronique sécurisée au Service statistique public, à des fins exclusives d'établissement de statistiques, les informations présentes dans les bases de données qu'elles détiennent, lorsque ces informations sont recherchées pour les besoins d'enquêtes statistiques obligatoires. Ces données ne peuvent pas ensuite être transmises à un tiers. » Pour accéder aux données privées suivant ce cadre, il est nécessaire de passer par plusieurs étapes définies dans le décret d'application : concertation avec les détenteurs de données, étude de faisabilité et d'opportunité, avis du Conseil national de l'information statistique (Cnis), parution d'un arrêté et surtout substitution d'une enquête statistique. **Les données de téléphonie mobile n'ont pour le moment pas été mobilisées pour obtenir des informations qui, sinon, auraient été collectées par une enquête. Elles ne sont donc a priori pas concernées par ce cadre légal**.

7 Voir les références juridiques en fin d'article.
8 Voir les références juridiques en fin d'article.

> ***Dans sa version révisée adoptée par le Parlement et le Conseil en 2024, le règlement européen n° 223/2009 introduit la possibilité légale d'utiliser les données détenues par le secteur privé pour le développement et la production de statistiques officielles européennes.***

L'utilisation des données de téléphonie mobile est encadrée, en plus du RGPD, par **le règlement européen n° 223/2009**[9], qui établit le cadre légal pour le développement, la production et la diffusion des statistiques européennes. Dans sa version révisée adoptée par le Parlement et le Conseil en 2024, il introduit la possibilité légale d'utiliser les données détenues par le secteur privé « pour le développement et la **production de statistiques officielles européennes**, sur une base durable et selon des règles équitables, claires, prévisibles et proportionnées, conformément au cadre des droits fondamentaux de l'Union. L'accès aux données détenues par le secteur privé devrait être garanti conformément au principe de rentabilité et ne doit pas entraîner de charge excessive pour les opérateurs économiques. » La question de la maille d'agrégation des données et du montant de l'éventuelle compensation financière des opérateurs reste ouverte, mais il s'agit d'une perspective prometteuse pour l'utilisation de ces données, sous réserve de la production de statistiques européennes rentrant dans le cadre du programme annuel de travail du système statistique européen. Pour mettre en place des collaborations constructives, le champ exact des données fournies aux instituts nationaux de statistique devra prendre en compte également les enjeux propres aux opérateurs.

… et prennent en compte les enjeux propres aux opérateurs pour établir des partenariats bénéfiques pour tous

Les opérateurs de téléphonie mobile sont intéressés par l'expertise technique des chercheurs et des statisticiens publics. En travaillant sur les données de téléphonie, qu'elles soient détaillées ou agrégées, ceux-ci soulèvent certains problèmes non encore résolus et mettent à l'épreuve les méthodes d'agrégation utilisées par l'opérateur, ce qui conduit à une amélioration générale de la qualité. De plus, les statisticiens publics ont accès à des données de référence qui peuvent aider à mieux calibrer les données des opérateurs privés, ou du moins à mettre en avant les limites des méthodes de calibrage actuelles.

Sur le plan de l'image publique, le point de vue des opérateurs est nuancé. D'un côté, ces derniers souhaitent maîtriser l'utilisation faite de données sensibles pour leurs utilisateurs, alors que les enjeux d'acceptabilité sociale des travaux menés sur les données sont cruciaux. D'un autre côté, travailler avec le milieu académique et la statistique publique illustre la contribution de l'opérateur aux sujets d'intérêt général, ce qui a un rôle positif sur son image (Sémécurbe et al., 2020 ; Coudin et al., 2021).

Les opérateurs craignent aussi la divulgation à leurs concurrents d'informations sensibles sur leurs parts de marché. Travailler avec les données détaillées anonymisées de plusieurs opérateurs nécessite donc de se positionner comme un tiers de confiance.

9 Voir les références juridiques en fin d'article.

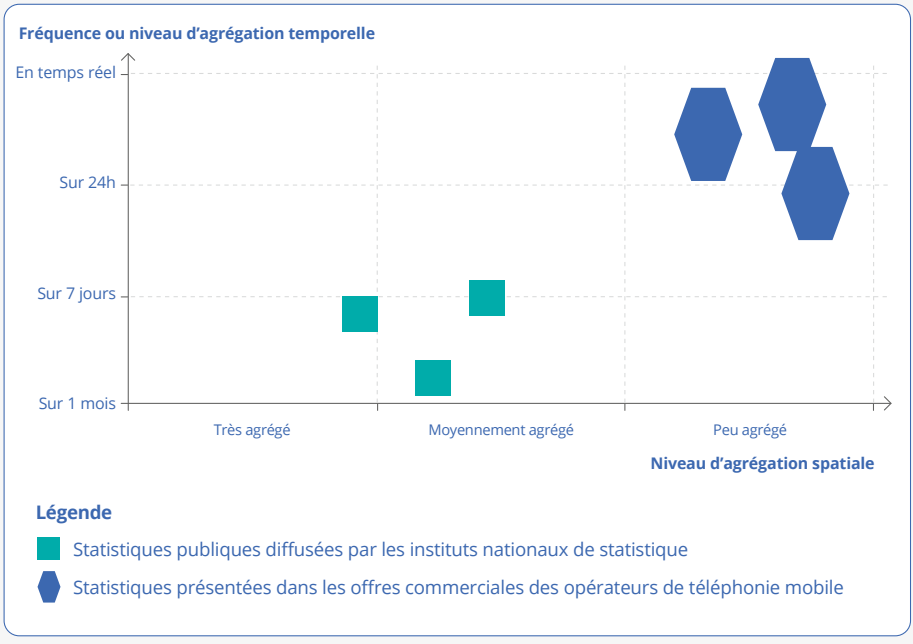
> ***Une bonne coopération avec les fournisseurs de données est un prérequis pour bien comprendre leur méthodologie de construction et produire des statistiques les plus robustes et documentées possible.***

Ceci demande un investissement non négligeable dans des infrastructures techniques permettant de garantir le niveau de confidentialité requis.

Enfin, certains opérateurs commercialisent eux-mêmes des statistiques agrégées, notamment pour étudier la population présente et les mobilités. Quel que soit le cadre législatif, il est nécessaire d'échanger avec eux pour établir un partenariat qui soit bénéfique pour toutes les parties, d'autant que ces données ne sont pas directement exploitables par les statisticiens et doivent être retraitées par les opérateurs. Une bonne coopération avec les fournisseurs de données est dans tous les cas un prérequis pour bien comprendre leur méthodologie de construction et produire des statistiques les plus robustes et documentées possible. En général, les indicateurs publiés par les instituts de statistique se situent à une maille géographique et temporelle beaucoup plus large que celle des indicateurs vendus par les opérateurs (***figure 3***).

Quels sont plus précisément les indicateurs pour lesquels l'utilisation des données de téléphonie mobile présente un intérêt pour les statistiques publiques ?

Figure 3 - Instituts nationaux de statistique et opérateurs de téléphonie : des niveaux d'analyse complémentaires*



* D'après une figure réalisée par Fabio Ricciato (Eurostat).

Une source d'information riche et utile pour le système statistique public

Grâce à leur volume et à la richesse des informations qu'elles contiennent, les données de téléphonie mobile peuvent compléter celles du système statistique public pour étoffer les statistiques habituellement produites ou pour éclairer un nouvel angle d'analyse.

Étudier la population présente

Le taux d'incidence de la Covid-19 a-t-il augmenté en Île-de-France durant les Jeux Olympiques de 2024 ? La réponse varie suivant le dénominateur : population résidente au sens du recensement, ou population réellement présente à cette période estimée avec la téléphonie mobile. En revenant sur la situation de 2021, Tarantola et Hamidouche (2025), chercheurs à Santé Publique France, montrent qu'en prenant au dénominateur la population présente, le critère d'alerte maximal a été dépassé durant le mois d'août 2021, alors que ce n'est pas le cas si on met au dénominateur la population résidente (***figure 4***).

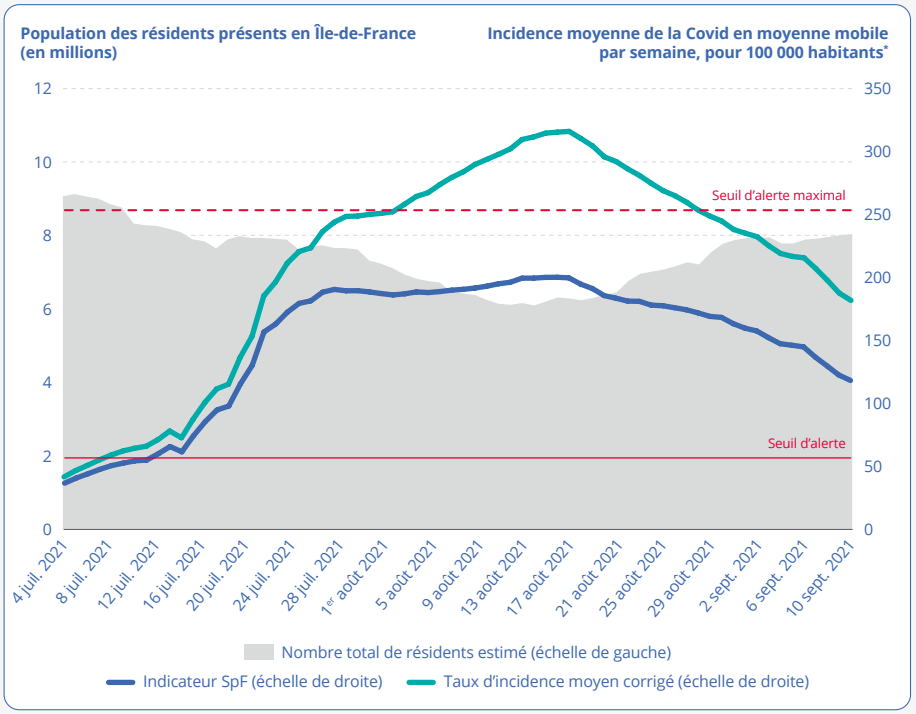
> ***De nombreuses politiques publiques gagneraient en précision et en efficacité en complétant les données de population résidente par celles de la population présente.***

Au-delà des politiques sanitaires, de nombreuses politiques publiques gagneraient en précision et en efficacité en complétant les données de population résidente par celles de la population présente. Ce constat a été fait également par le Cnis. Ce dernier indique dans ses objectifs de moyen terme 2024-2028 que « la seule mesure de la population résidente ne suffit pas pour capter la dynamique et l'attractivité d'un territoire, la fréquentation de ses équipements et l'utilisation de ses ressources ». Il émet le souhait que « l'ensemble des données publiques et privées soient mobilisées pour mesurer la population présente ». Insistons bien, malgré tout, sur le fait que la « population présente » est un autre concept que celui de « population résidente » et que seul le recensement de la population permet d'estimer cette dernière.

Pour estimer la population présente, l'étape de mise en cohérence (calage) des données de téléphonie mobile avec les données de référence est particulièrement importante. Pour ce faire, l'enjeu méthodologique décrit plus haut d'une bonne attribution du lieu de domicile est particulièrement important. Une fois les données calées, il est envisageable de les confronter à d'autres sources de données issues d'acteurs privés, de façon à aborder l'évolution de la population présente sous différents angles. Ainsi, la ***figure 5*** compare le nombre de personnes détectées avec la téléphonie mobile au nombre de personnes ayant réalisé une transaction par carte bancaire, demi-heure par demi-heure, sur la zone de Lyon en septembre 2022[10]. Les données de téléphonie mobile correspondent à la moyenne du nombre de personnes ayant borné dans un Iris (quartier) de la ville de Lyon. Ces données ont été mises à disposition par le service Flux Vision commercialisé par Orange Business, de façon agrégée. Il s'agit davantage d'un

10 Source : Gabrielle Gambuli et Chloé Breton (Insee, université Gustave Eiffel, Télécom Paris).

Figure 4 - Incidence de la Covid-19 pendant l'été 2021, par semaine glissante, calculée par Santé publique France (SpF) et corrigée



* Au numérateur : nombre de tests positifs réalisés par les résidents d'Île-de-France en Île-de-France ; au dénominateur, nombre de résidents d'Île-de-France au sens de l'Insee pour l'indicateur SpF ou estimés avec les données de téléphonie pour l'indicateur corrigé.
Champ : Île-de-France.
Sources : Santé publique France ; Orange (données de signal) ; calculs Santé publique France.
D'après Tarantola et Hamidouche, à paraître.

indicateur de fréquentation que d'un indicateur de présence. En effet, une personne qui traverse plusieurs fois un Iris durant la demi-heure est comptabilisée plusieurs fois. Les données de transaction par carte bancaire correspondent au nombre de porteurs de cartes différents ayant effectué une transaction dans la zone durant la demi-heure. L'accès aux données individuelles accordé aux chercheurs dans le cadre de la chaire Finance digitale[11] a permis de construire un indicateur de fréquentation calculé avec une méthodologie identique à celle utilisée pour les données de téléphonie mobile.

En semaine, on observe avec les données de téléphonie mobile une augmentation de fréquentation de l'aire d'attraction de la ville de Lyon en journée, avec un premier pic lors des déplacements domicile-travail matinaux, un deuxième à la pause déjeuner et un troisième, plus important, au moment du retour à domicile. Les volumes de cartes bancaires ayant effectué une transaction, eux, augmentent nettement au moment de la pause déjeuner et en fin d'après-midi. Le samedi, les augmentations de fréquentation observées grâce

11 Voir l'article de Boittelle et al. sur les données de transactions par carte bancaire CB dans ce même numéro.

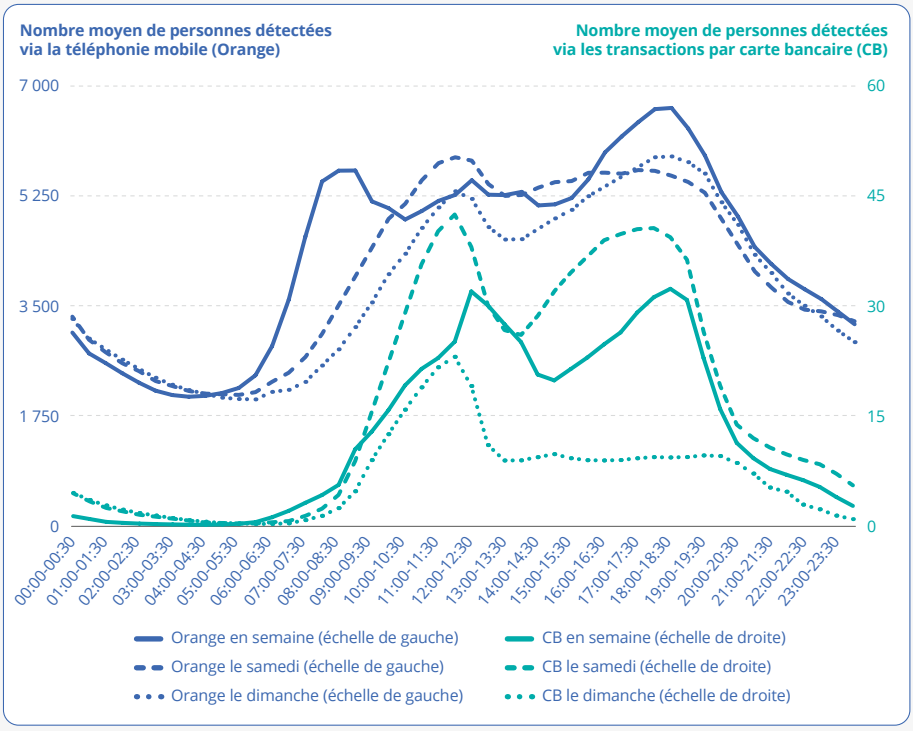
> ***Ces travaux exploratoires illustrent l'intérêt d'utiliser plusieurs sources de données complémentaires pour affiner l'analyse des comportements de la population.***

aux données de téléphonie mobile (en fin de matinée et dans l'après-midi) coïncident cette fois avec l'augmentation des transactions par carte bancaire. Enfin, le dimanche, les transactions par carte bancaire sont plus faibles, en particulier l'après-midi. Ces travaux exploratoires illustrent l'intérêt d'utiliser plusieurs sources de données complémentaires pour affiner l'analyse des comportements de la population. Combiner ces visions parcellaires de la situation réelle permet un vrai gain en qualité.

Aller au-delà des mobilités domicile-travail

L'Insee publie tous les dix ans différents zonages d'étude, destinés à mieux comprendre l'organisation du territoire et faciliter la production de statistiques territoriales. Ces zonages s'appuient souvent sur les déplacements domicile-travail mesurés par le

▶ Figure 5 - Volume moyen de personnes détectées dans l'aire d'attraction de la ville de Lyon, avec la téléphonie mobile et les transactions par carte bancaire, en septembre 2022



Sources : Flux Vision (données de signal) ; groupement des cartes bancaires CB ; calculs Insee.
D'après Gabrielle Gambuli et Chloé Breton (Insee, université Gustave Eiffel, Télécom Paris).

recensement. Cette source a l'avantage d'être exhaustive, fiable et bien documentée, mais les déplacements mesurés ne couvrent qu'une partie des mobilités : les déplacements des non-actifs (retraités, étudiants, etc.) façonnent aussi le territoire. Depuis la crise de la Covid-19, par ailleurs, le développement du télétravail a fait évoluer les mobilités quotidiennes des actifs. Par exemple, le zonage en aires d'attraction des villes[12] définit les pôles, densément peuplés et riches en emplois, puis leurs couronnes, communes dont 15 % des actifs occupés vont travailler dans le pôle. Les contours des aires d'attraction des villes évolueront sûrement si l'on considère l'ensemble des déplacements observés dans les données de téléphonie mobile et pas uniquement les déplacements domicile-travail.

Les mobilités sont en partie captées grâce à des enquêtes spécifiques (par exemple les enquêtes EMC2[13] du Cerema[14] ou l'enquête « Mobilité des personnes »[15] du SDES[16]). Comme indiqué dans le suivi d'avis du Cnis du 6 juin 2024[17], les données d'enquête sont riches d'informations concernant le profil des personnes présentes et leurs motifs de présence, mais elles ne permettent pas d'observer les variations hebdomadaires ou mensuelles et certaines ne couvrent que les espaces urbains. Ces enquêtes et les données de téléphonie mobile interviennent donc de façon complémentaire.

Analyser l'évolution de la ségrégation sociospatiale au fil de la journée

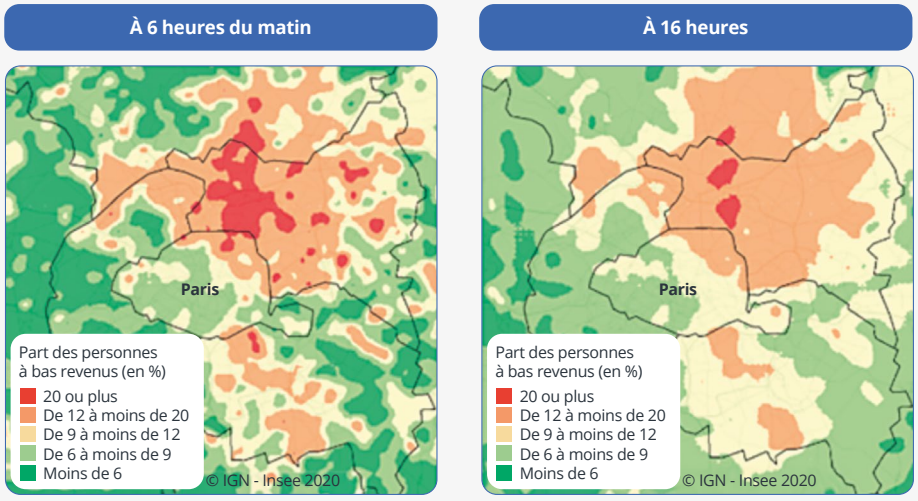
Les quartiers prioritaires de la politique de la ville concentrent de nombreuses difficultés sociales et économiques. Mieux comprendre dans quelle mesure ces quartiers sont isolés du reste des espaces urbains est un fort enjeu pour cibler les politiques publiques visant à diminuer cet isolement. Les indicateurs de mixité sociale pendant la nuit permettent d'identifier les zones où la ségrégation sociale sur le lieu de résidence est la plus forte. Toutefois, en fonction des mobilités en journée, cette ségrégation peut rester très marquée dans certains espaces et au contraire s'atténuer dans d'autres.

Étudier ce phénomène nécessite de combiner données de téléphonie mobile et données traditionnelles à l'échelle spatiale la plus fine possible. Bien sûr, l'appariement exact au niveau individuel n'est pas envisageable puisque toutes les données sont anonymisées. Il est toutefois possible d'attribuer à tous les porteurs de téléphone mobile résidant dans une zone géographique donnée (commune, carreau[18], etc.) les caractéristiques sociodémographiques des habitants de cette zone. La mobilité moyenne observée en journée pour les habitants de cette zone permet ensuite d'étudier leur nouvelle répartition sur le territoire heure par heure et ainsi l'évolution de la ségrégation.

À 6 heures du matin, on considère que les personnes détectées dans les données de téléphonie mobile sont à leur lieu de domicile. À cette heure, en région parisienne, les zones où la part des personnes à bas revenus est supérieure à 20 % sont très concentrées dans le

12 *https://www.insee.fr/fr/metadonnees/definition/c2173.*
13 *https://www.cerema.fr/fr/activites/mobilites/connaissance-modelisation-evaluation-mobilite/enquetes-mobilite-emc2.*
14 Le Cerema est le Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement.
15 *https://www.statistiques.developpement-durable.gouv.fr/enquete-sur-la-mobilite-des-personnes-2018-2019?list-enquete=true.*
16 Le service des données et études statistiques (SDES) assure les fonctions de service statistique des ministères chargés de l'environnement, de l'énergie, de la construction, du logement et des transports.
17 *https://www.cnis.fr/wp-content/uploads/2024/03/com-terr-2024-1-suivi-avis.pdf.*
18 L'Insee découpe le territoire en carreaux pour y diffuser de l'information statistique à un niveau faiblement agrégé. Selon les informations diffusées, il s'agit de carreaux de 1 km de côté ou, au plus fin,de 200 mètres de côté. Voir *https://www.insee.fr/fr/outil-interactif/7737357/documentation.html#carroyage.*

▶ Figure 6 - Évolution de la ségrégation sociospatiale en journée



Lecture : Les personnes à faibles revenus sont plus concentrées dans le nord-est durant la nuit, la ségrégation diminue au cours de la journée.
Champ : Unité urbaine de Paris.
Sources : Orange, CDR 2007 ; Insee, Filosofi 2014 ; calculs Insee. D'après les figures 1.b et 1.c publiées dans Galiana et al. (2020).

nord de Paris, notamment en Seine-Saint-Denis (***figure 6***). En revanche, à 16h, la répartition est beaucoup plus homogène. Les données de téléphonie mobile permettent ainsi d'éclairer l'évolution de la ségrégation sociospatiale en journée d'une façon que ne permettent pas du tout les sources traditionnelles des statistiques publiques (Galiana et al., 2020).

▶ Des perspectives européennes

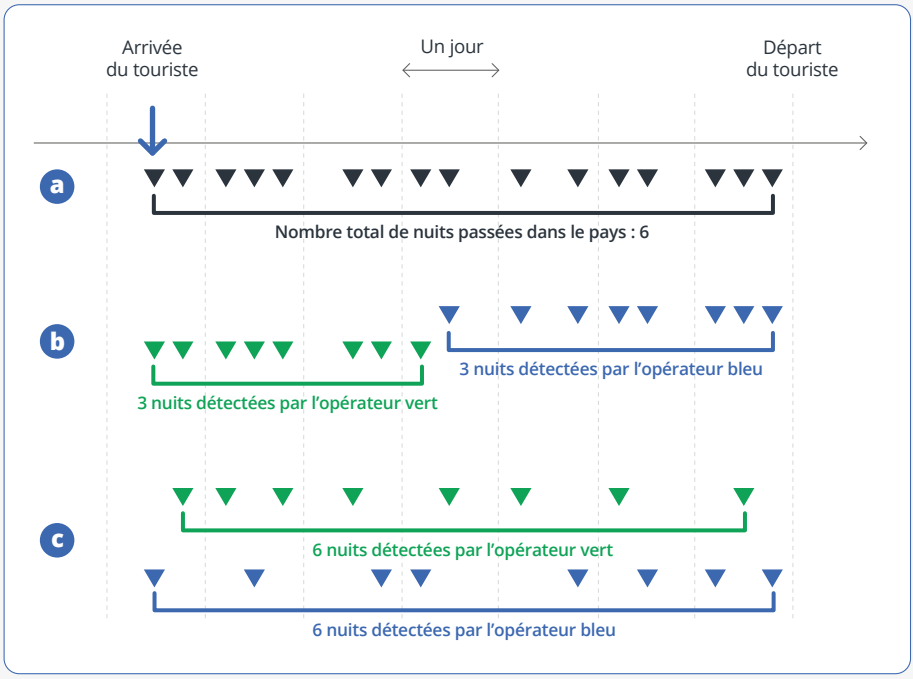
L'enjeu de la prise en compte de l'itinérance

> ***Au-delà de la France, l'usage des données de téléphonie mobile est un enjeu pour l'ensemble du système statistique public européen.***

Au-delà de la France, l'usage des données de téléphonie mobile est un enjeu pour l'ensemble du système statistique public européen (*European Statistical System*, 2021). Certaines thématiques dépassent par définition les frontières nationales. Ainsi, étudier les mobilités touristiques nécessite de traiter au mieux le fait que l'abonné d'un opérateur peut en changer lorsqu'il visite un autre pays (phénomène d'itinérance ou *roaming*) et même plusieurs fois. Ces changements peuvent advenir y compris lorsque l'abonné est dans son propre pays mais sur une zone frontalière. Le fait que les opérateurs ne collaborent pas au niveau des données individuelles peut mener à des doubles comptes. Ainsi, la ***figure 7*** présente le cas d'un

touriste passant six nuits en France. Dans le cas (b), l'opérateur vert détecte trois nuits, puis l'opérateur bleu détecte trois nuits. Il n'y a donc pas de doubles comptes, mais si l'institut de statistique a accès aux données d'un unique opérateur, le visiteur ne sera observé que partiellement. En revanche, dans le cas (c), les opérateurs vert et bleu détectent chacun six nuits. Si les deux opérateurs fournissent des décomptes agrégés, le visiteur sera compté en double. Il s'agit là des principaux cas, mais de nombreuses autres configurations peuvent exister. Une bonne collaboration entre opérateurs et entre instituts de statistique au-delà des frontières nationales est ainsi fondamentale pour estimer les flux avec une précision suffisante.

Figure 7 - Quelques conséquences possibles du phénomène d'itinérance



Lecture : Un touriste passe 6 nuits en France. Dans le cas (b), chacun des deux opérateurs détecte trois nuits. Dans le cas (c), les opérateurs détectent chacun six nuits. Le triangle représente le signal émis par le téléphone.
Source : projet Multi-MNO : *https://cros.ec.europa.eu/landing-page/multi-mno-project*.

De nombreux projets en cours sous l'égide d'Eurostat

Eurostat finance depuis décembre 2023 et pour une durée de deux ans, le projet de recherche « *Trusted Smart Statistics: methodological developments based on new data sources* »[19], appelé également projet MNO-MINDS[20]. Ce projet regroupe dix pays européens et est coordonné par l'Italie. L'objectif est de développer un cadre méthodologique commun permettant de combiner données de téléphonie mobile et autres sources de données. Les données considérées appartiennent à deux grandes catégories : celles collectées par les instituts nationaux de statistique (INS) dans le but d'élaborer des statistiques (recensement, enquêtes) ; celles collectées en premier lieu pour des usages autres que statistiques et ensuite réutilisées pour les analyses socioéconomiques (données administratives, capteurs de trafic routier, données de billettique de transport en commun, etc.). Eurostat souhaite que ce panorama des sources prenne en compte l'arbitrage entre la qualité des données et leur coût (d'acquisition ou de traitement suivant la source). L'objectif est de se placer dans une perspective de production régulière de statistiques officielles et plus seulement de statistiques expérimentales. De plus, il s'agit de considérer la disponibilité potentielle dans tous les pays européens. Des méthodes permettant de combiner les différentes sources de données seront également développées et diffusées et une enquête ad hoc servira à mieux comprendre les usages des téléphones : distinction entre téléphones personnel et professionnel, cas où un abonné principal gère les téléphones de toute sa famille, etc.

> ***L'objectif est de se placer dans une perspective de production régulière de statistiques officielles et plus seulement de statistiques expérimentales.***

Par ailleurs, Eurostat finance un deuxième projet (Multi-MNO[21]), débuté en décembre 2023 pour une durée de deux ans, piloté par un cabinet de conseil (GOPA) et réunissant des INS de quatre pays (CBS pour les Pays-Bas, ISTAT pour l'Italie, GUS pour la Pologne et SURS pour la Slovénie), des entreprises spécialisées dans le traitement de données mobiles (Positium et Nommon) et cinq opérateurs de téléphonie mobile, issus de quatre pays différents (Orange Espagne, Vodafone Espagne, Vodafone Italie, A1 Slovénie et POST Luxembourg). L'objectif est de développer un traitement de données standardisé (ou *pipeline*) permettant d'agréger les données individuelles des opérateurs sur la base d'une méthode validée, de façon à produire des statistiques agrégées avec un bon niveau de qualité.

Le fait que deux opérateurs du même pays participent au projet permettra de démontrer, d'un point de vue technique et organisationnel, la possibilité de produire des statistiques à partir de données d'opérateurs concurrents au sein d'un même pays. Les opérateurs fourniront l'accès à leurs données pour tester, évaluer et améliorer le circuit de traitement développé durant le projet. Seules les données agrégées et anonymisées quitteront

19 Des statistiques de confiance à partir des objets connectés : développements méthodologiques basés sur les nouvelles sources de données. *https://www.insee.fr/fr/information/7681963*.
20 *https://cros.ec.europa.eu/mno-minds*.
21 *https://cros.ec.europa.eu/landing-page/multi-mno-project*.

> ***Une task force sur l'usage des données de téléphonie mobile pour la statistique publique a été lancée par Eurostat en 2021.***

les serveurs des opérateurs de téléphonie mobile. L'objectif est que les traitements sur données individuelles développés dans le cadre du projet puissent être réalisés directement sur les serveurs des opérateurs. Certains éléments de la méthode d'agrégation resteront paramétrables, de façon à s'ajuster aux exigences nationales et notamment aux contraintes réglementaires, lesquelles peuvent différer suivant les pays. Une attention particulière sera portée au respect des informations commerciales sensibles pour les opérateurs. Par exemple, les poids utilisés pour garantir la représentativité des statistiques ne seront pas publiés. Des experts du domaine juridique font partie du consortium.

Enfin, une task force sur l'usage des données de téléphonie mobile pour la statistique publique a été lancée par Eurostat en 2021. Son objectif est de coordonner et d'orienter les développements méthodologiques relatifs à l'utilisation de données de téléphonie mobile au sein du système statistique européen, de favoriser le partage des connaissances et les bonnes pratiques issues des expériences nationales, et de progresser vers la définition d'un cadre méthodologique commun pour l'ensemble du système statistique européen.

En septembre 2023, les 18 pays membres de la task force ont cosigné un article (*European Statistical System Task Force*, 2023) argumentant en faveur du développement d'une méthodologie commune à l'ensemble du système statistique européen pour traiter les données de téléphonie mobile. Une telle méthodologie doit être transparente et interprétable. Elle doit également permettre de comparer les statistiques obtenues dans les différents pays et de combiner différentes statistiques entre elles.

Remerciements

Pierre Bayard, Chloé Breton, Étienne Côme, Élise Coudin, Gabrielle Gambuli, Mélina Hillion, Sylvie Lagarde, Arnaud Legendre, Romain Lesur, Fanny Mikol, Latifa Oukhellou, Corinne Prost, Patrick Redor, Denis Renaud, Milena Suarez Castillo. Remerciement au projet MobiTIC (ANR-19-CE22-0010), financé par l'Agence nationale de la recherche en France.

▶ Fondements juridiques

- Règlement (CE) n° 223/2009 révisé du Parlement européen et du Conseil du 11 mars 2009 relatif aux statistiques européennes. In : *site de l'Union européenne*. [en ligne]. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/?qid=1724682778358&uri=CELEX%3A32009R0223*.

- Règlement (UE) n° 2022/868 du Parlement européen et du Conseil du 30 mai 2022 portant sur la gouvernance européenne des données et modifiant le règlement (UE) n° 2018/1724. In : *site de l'Union européenne*. [en ligne]. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://eur-lex.europa.eu/eli/reg/2022/868/oj?locale=fr*.

- Loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques. In : *site de Légifrance*. [en ligne]. Mise à jour le 25 mars 2019. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573*.

- Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique. In : *site de Légifrance*. [en ligne]. [Consulté le 25 mars 2025]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746*.

▶ Bibliographie

- ARCEP, 2024. *Mon réseau mobile*. [en ligne]. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://monreseaumobile.arcep.fr/*.

- AUDE, Johanne, DEBOUZY, Ivan, JOUBERT, Marie-Pierre, PRAMIL, Julien et GAMBULI, Gabrielle, 2024. Cinq types de territoires diversement habités et inégalement fréquentés en journée – Aire d'attraction de la ville de Lyon. In : *Insee Analyses*. [en ligne]. 27 novembre 2024. Insee. N° 184. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/8289238*.

- BONNETAIN, Loïc, FURNO, Angelo, EL FAOUZI, Nour-Eddin, FIORE, Marco, STANICA, Razvan, SMOREDA, Zbigniew et ZIEMLICKI, Cezary, 2021. TRANSIT: Fine-grained human mobility trajectory inference at scale with mobile network signaling data. In : *Transportation Research Part C: Emerging Technologies*. [en ligne]. Septembre 2021. Volume 130, 103257. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://doi.org/10.1016/j.trc.2021.103257*.

- COMBES, Stéphanie, JOUBERT-DE BELLEFON, Marie-Pierre et VANHOOF, Maarten, 2017. Mining Mobile Phone Data to Detect Urban Areas. In : *Proceedings of the Conference of the Italian Statistical Society: SIS 2017, Statistics and Data Science: New challenges, new generations*. [en ligne]. Juin 2017. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://www.researchgate.net/publication/321278660_Mining_Mobile_Phone_Data_to_Detect_Urban_Areas*.

- COUDIN, Élise, POULHES, Mathilde et SUAREZ CASTILLO, Milena, 2021. The French official statistics strategy: Combining signaling data from various mobile network operators for documenting COVID-19 crisis effects on population movements and economic outlook. In : *Cambridge University Press*. [en ligne]. 24 juin 2021. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://doi.org/10.1017/dap.2021.1*.

- EUROPEAN STATISTICAL SYSTEM, 2021. European Statistical System (ESS) position paper on the future Data Act proposal – Access to privately held data is urgently needed for producing new, faster, more detailed official statistics. In : *site d'Eurostat*. [en ligne]. Juin 2021. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/documents/13019146/13405116/main+ESS+position+paper+on+future+Data+Act+proposal.pdf/37f3b5c7-abfd-5a05-6be2-fdc4b87ee7d2?t=1631695372906*.

- EUROPEAN STATISTICAL SYSTEM TASK FORCE, 2023. Reusing mobile network operator data for official statistics: the case for a common methodological framework for the European Statistical System. In : *site de Eurostat*. [en ligne]. 12 septembre 2023. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/web/products-statistical-reports/w/ks-ft-23-001*.

- GALIANA, Lino, SUAREZ CASTILLO, Milena, SÉMÉCURBE, François, COUDIN, Élise et JOUBERT-DE BELLEFON, Marie-Pierre, 2020. Retour partiel des mouvements de population avec le déconfinement. In : *Insee Analyses*. [en ligne]. 22 juillet 2020. Insee. N° 54. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4635407*.

- GALIANA, Lino, SAKAROVITCH, Benjamin, SÉMÉCURBE, François et SMOREDA, Zbigniew, 2020. La mixité sociale est plus forte en journée sur les lieux d'activité que pendant la nuit dans les quartiers de résidence. In : *Insee Analyses*. [en ligne]. 9 novembre 2020. Insee. N° 59. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4930403*.

- GOOTZEN, Yvonne et TENNEKES, Martijn, 2022. Bayesian location estimation of mobile devices using a signal strength model. In : *Journal of Spatial Information Science*. [en ligne]. Décembre 2022. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://josis.org/index.php/josis/article/view/166*.

- RICCIATO, Fabio et COLUCCIA, Angelo, 2021. On the Estimation of Spatial Density from Mobile Network Operator Data. In : *IEEE Transactions on Mobile Computing*. [en ligne]. Décembre 2021. Pp. (99):1-1. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://www.researchgate.net/publication/357009792_On_the_estimation_of_spatial_density_from_mobile_network_operator_data*.

- RICCIATO, Fabio, LANZIERI, Giampaolo, WIRTHMANN, Albrecht et SEYNAEVE, Gerdy, 2020. Towards a methodological framework for estimating present population density from mobile network operator data. In : *Pervasive and Mobile Computing*. [en ligne]. Octobre 2020. Volume 68. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://doi.org/10.1016/j.pmcj.2020.101263*.

- SAKAROVITCH, Benjamin, JOUBERT-DE BELLEFON, Marie-Pierre, GIVORD, Pauline et VANHOOF, Maarten, 2019. Estimer la population résidente à partir de données de téléphonie mobile, une première exploration. In : *Économie et Statistique / Economics and Statistics*. [en ligne]. 11 avril 2019. N° 505-506. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/3706213?sommaire=3706255*.

- SALGADO, David, SANGUIAO, Luis, OANCEA, Bogdan, BARRAGÁN, Sandra et NECULA, Marian, 2021. An end-to-end statistical process with mobile network data for official statistics. In : *EPJ Data Science*. [en ligne]. 29 avril 2021. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-021-00275-w*.

- SÉMÉCURBE, François, SUAREZ CASTILLO, Milena, GALIANA, Lino, COUDIN, Élise et POULHES, Mathilde, 2020. Que peut faire l'Insee à partir des données de téléphonie mobile ? – Mesure de population présente en temps de confinement et statistiques expérimentale. In : *Le blog de l'Insee*. [en ligne]. 15 avril 2020. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://blog.insee.fr/que-peut-faire-linsee-a-partir-des-donnees-de-telephonie-mobile-mesure-de-population-presente-en-temps-de-confinement-et-statistiques-experimentales/*.

- SUAREZ CASTILLO, Milena, SÉMÉCURBE, François, ZIEMLICKI, Cezary, TAO, Haixuan X. et SEIMANDI, Tom, 2023. Temporally Consistent Present Population from Mobile Network Signaling Data for Official Statistics. In : *Journal of Official Statistics*. [en ligne]. 10 décembre 2023. Vol. 39, N° 4, pp. 535–570. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://sciendo.com/article/10.2478/jos-2023-0025*.

- TARANTOLA, Arnaud et HAMIDOUCHE, Mohamed, 2025. Cell phone data to correct population estimates and Sars-Cov2 incidence in Ile-de-France during the summer of 2021. In : *Eurosurveillance*. À paraître.

- VANHOOF, Maarten, REIS, Fernando, PLOETZ, Thomas et SMOREDA, Zbigniew, 2018. Assessing the Quality of Home Detection from Mobile Phone Data for Official Statistics. In : *Journal of Official Statistics*. [en ligne]. Décembre 2018. Vol. 34, N° 4, pp. 935–960. [Consulté le 27 décembre 2024]. Disponible à l'adresse : *https://www.researchgate.net/publication/329712197_Assessing_the_Quality_of_Home_Detection_from_Mobile_Phone_Data_for_Official_Statistics*.

Les données de transactions par carte bancaire CB

Quels apports possibles aux analyses conjoncturelles et territoriales ?

Mathieu Boittelle *, *Émilie Cupillard***, *Alain Jacquot****
*Marie-Pierre Joubert***** *et Florian Le Goff******

Depuis le printemps 2020, l'Insee dispose, dans des délais très courts, d'agrégats de paiement par carte bancaire par secteur d'activité. Ces derniers lui sont transmis par le schéma (ou réseau) domestique français de paiement par carte et mobile « CB », principal réseau utilisé sur le territoire. L'Insee utilise ces agrégats pour produire une estimation avancée du volume des ventes dans le commerce de détail. L'institut a également mobilisé ces agrégats pour mesurer l'impact sur la consommation des confinements imposés pendant la crise sanitaire.

Complétées par d'autres sources, ces données sont ainsi intéressantes pour produire des premières estimations de la consommation et de l'activité dans certains secteurs tels que le commerce de détail. Cependant, la couverture de la consommation par les paiements par carte bancaire CB fluctue au cours du temps en raison de l'existence de moyens de paiement alternatifs (espèces, chèques, transactions par carte via un réseau de paiement international comme Visa ou Mastercard, etc.), mais aussi d'évolutions des comportements. De ce fait, les sources traditionnelles restent irremplaçables pour produire les estimations définitives.

Dans le cadre de la chaire Finance digitale, les membres habilités des programmes de recherche CB-Insee mobilisent par ailleurs, de manière sécurisée et anonymisée, des données détaillées de paiement par carte bancaire CB. Ces dernières constituent une source prometteuse pour éclairer, dans le cadre de ces travaux de recherche, des problématiques telles que les connexions entre commerces et territoires.

Since the spring of 2020, INSEE has had access in a very short time-span to aggregates of bankcard payments, by sector of activity. This data is transmitted to the institute by the French domestic mobile and bankcard payment "CB" scheme (or network), the most prominent payment operator in the French territory. INSEE uses these aggregates to produce an early estimate of sales volume in the retail trade. The institute also mobilised them to quantify the impact on consumption of the lockdowns imposed during the health crisis.

Supplemented by other sources, this data is therethore interesting for producing early estimates of consumption and activity in some sectors, as in retail trade. However, the coverage of consumption by CB bankcard payments fluctuates over time due to the existence of alternative means of payment (cash, cheques, cards issued by other payment networks such as Visa or Mastercard, etc.) and changes in behaviour. As a result, traditional sources are still essential for computing definitive estimates of these indicators.

As part of the Finance Digitale chair, the authorised members of the CB/INSEE research programmes also mobilise, in a secure and anonymous way, disaggregated bankcard payment data. This is a promising source for shedding light, within the confines of the research projects, on issues such as connections between shops and areas.

* Chargé d'études, Direction régionale d'Occitanie, Insee. *mathieu.boittelle@insee.fr*

** Cheffe de la section Consommation des ménages et innovations méthodologiques, DESE, Insee. *emilie.cupillard@insee.fr*

*** Directeur de projet, Département de la conjoncture, DESE, Insee. *alain.jacquot@insee.fr*

**** À la date de rédaction de l'article, responsable adjointe du SSP Lab, DMCSI, Insee. *marie-pierre.joubert@finances.gouv.fr*

***** À la date de rédaction de l'article, expert en méthodologie statistique, division Indicateurs conjoncturels d'activité, DSE, Insee. *florian.le-goff@insee.fr*

Au sein des économies développées, la carte bancaire est devenue un moyen de paiement très courant pour les particuliers, que ce soit pour des achats en magasin ou en ligne. En France, dans les magasins, son usage dépasse pour la première fois en 2024 celui des espèces en nombre de transactions : elle est mobilisée dans 48 % des transactions contre 43 % pour les espèces (Banque de France, 2025). Pour l'ensemble des achats en magasin et en ligne, en dehors des paiements en espèce, la carte concernait par ailleurs 61 % des transactions réalisées par les particuliers, les entreprises et les administrations (Observatoire de la sécurité des moyens de paiement, 2024). Cette part n'a cessé de croître au fil des ans, à l'inverse en particulier de celle des chèques (***figure 1***).

▶ Des données d'origine privée pouvant enrichir les statistiques publiques…

Une transaction de paiement par carte fait intervenir sept acteurs :

- l'acheteur (le porteur de la carte) et sa banque (l'émetteur de la carte),
- le commerçant et sa banque (dite banque acquéreur),
- les acteurs qui organisent la transaction : le schéma de paiement (CB, Visa, Mastercard, etc.), qui pilote cette organisation, ainsi que, sur le plan opérationnel, le réseau interbancaire d'autorisation et le système interbancaire de compensation (*clearing* en anglais).

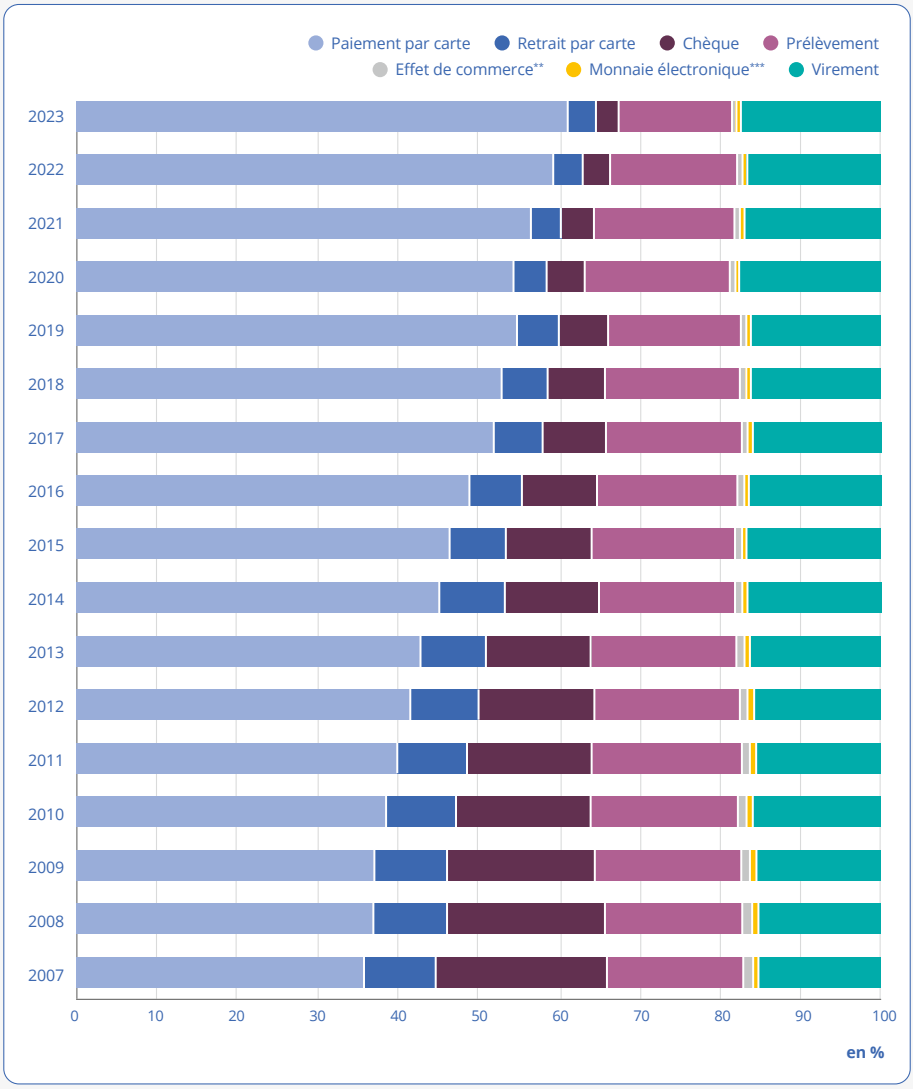
Le **schéma de paiement** définit notamment les règles, les exigences techniques, les mesures de sécurité et met en œuvre les outils interbancaires de lutte contre la fraude nécessaires au bon fonctionnement et à la sécurité du schéma de paiement. En France, les transactions passent essentiellement par le schéma de paiement domestique français CB, piloté par le groupement d'intérêt économique Cartes Bancaires CB ou GIE CB (***encadré***).

Le **réseau interbancaire d'autorisation** (e-rsb, opéré par la société STET, pour le schéma CB) achemine jusqu'à l'émetteur de la carte les demandes d'autorisation émises par les terminaux de paiement électroniques ou les sites internet. Cela permet à l'émetteur d'effectuer différents contrôles avant d'autoriser un paiement ou un retrait par carte. Il s'agit par exemple de vérifier que le plafond de paiement de la carte n'est pas dépassé ou que la carte n'est pas en opposition.

Le **système interbancaire de compensation** (CORE, opéré par la société STET, pour le schéma CB) organise l'échange financier de la transaction de paiement entre la banque du commerçant et celle du porteur de carte. Cela permet in fine de débiter le compte bancaire du porteur et de créditer le compte bancaire du commerçant, dans leurs banques respectives, du montant de la transaction (***figure 2***).

L'usage de plus en plus répandu de la carte conduit à traiter et enregistrer de nombreuses données sur les consommateurs et les commerces dans les systèmes d'information des banques, des schémas de paiement et des opérateurs associés. Une transaction par carte génère en particulier les données suivantes : le numéro de la carte du porteur, le montant et l'horodatage de la transaction, la nature de la transaction (paiement de

▶ Figure 1 - Répartition annuelle des moyens de paiement hors espèces, de 2007 à 2023, en volume*

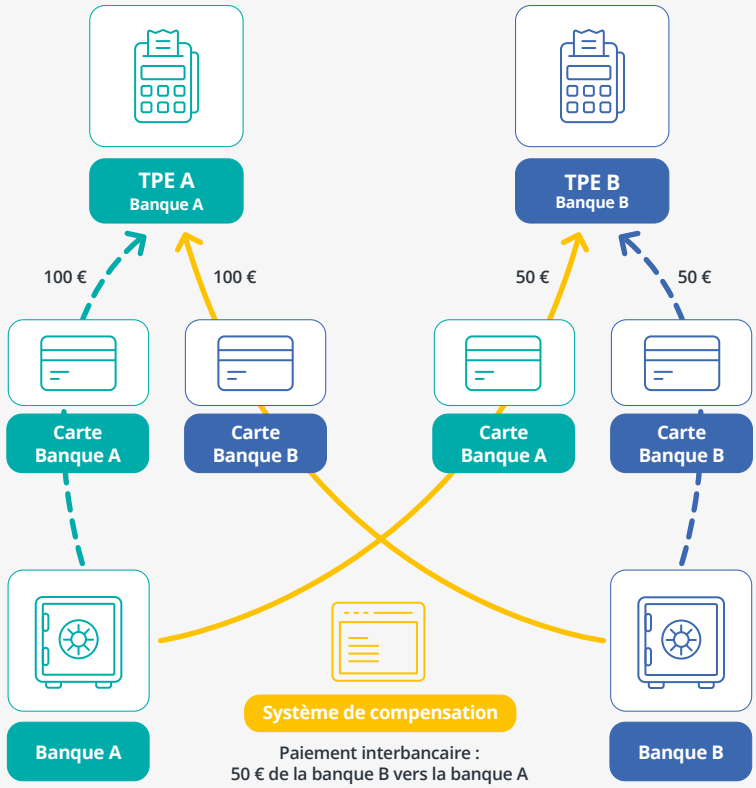


* Part dans le nombre total de transactions.
** Moyen de paiement non rattaché à un organisme bancaire, propre aux entreprises.
*** Paiements réalisés au travers d'un instrument préchargé en euros, sur un support physique (par exemple, carte prépayée) ou avec un portefeuille en ligne (par exemple, PayPal ou Lydia).
Lecture : En France, en 2023, 61 % des transactions ont été réalisées avec un paiement par carte, contre 36 % en 2007.
Champ : France, ensemble des transactions réalisées par les particuliers, les entreprises et les administrations.
Source : Observatoire de la sécurité des moyens de paiement.

proximité ou paiement en ligne), sa localisation géographique, ainsi que l'identification du commerçant (enseigne ou raison sociale, numéro Siret[1]). Ces informations sont enregistrées par les banques respectives du porteur et du commerçant. Elles transitent par le réseau interbancaire d'autorisation et le système interbancaire de compensation lorsque la banque du commerçant diffère de celle de l'acheteur.

Les données issues de ces actes de gestion (***figure 3***), à haute fréquence, produites dans des délais très courts et géolocalisées, sont susceptibles d'intéresser le statisticien public. Elles permettent d'appréhender certains phénomènes économiques dès lors qu'elles sont agrégées selon les dimensions pertinentes. Complétées par le domaine d'activité du commerçant, elles peuvent contribuer à améliorer les indicateurs conjoncturels

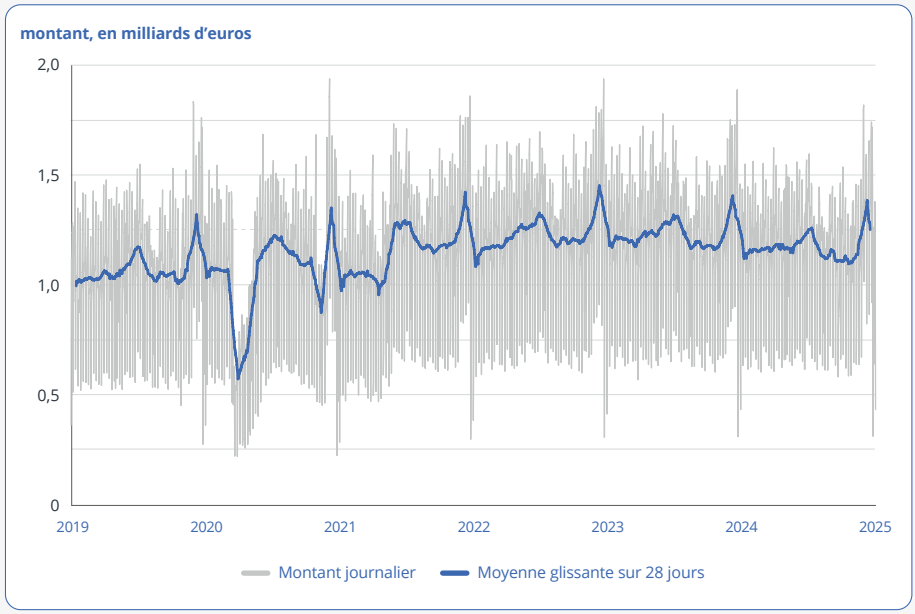
▶ Figure 2 - Gestion de la compensation interbancaire par le système interbancaire de compensation



TPE : Terminal de paiement électronique.
Lecture : Le schéma représente quatre transactions par carte bancaire. Dans deux d'entre elles (flèches en pointillés), la banque du client (porteur de carte) est aussi celle du commerçant (TPE). Dans les deux autres (flèches jaunes), les deux banques diffèrent : les transactions passent alors par le système de compensation, qui verse ensuite le solde de la banque débitrice à la banque créditrice.

1 *https://www.insee.fr/fr/metadonnees/definition/c1841.*

▶ Figure 3 - Montant journalier des transactions par carte bancaire CB de 2019 à 2024



Source : Cartes Bancaires CB, données de compensation pour les paiements par carte CB ; calculs Insee.

d'activité du commerce de détail, faisant gagner en qualité et en précocité. Enrichies de la localisation du commerçant, elles permettent de compléter la connaissance sur le commerce et les comportements de consommation, mais aussi d'évaluer les effets de chocs externes sur un territoire et, par le biais des trajectoires de paiement des cartes, d'appréhender les connexions entre territoires.

▶ … accessibles par l'Insee grâce à un cadre juridique spécifique à la France

À la fin des années 2010, l'Insee s'est mis à la recherche de nouvelles sources de données pour améliorer l'estimation précoce (à +30 jours) des indices de production et de chiffres d'affaires dans le commerce de détail (Cazaubiel et al., 2022). L'Insee est tenu de produire une telle estimation précoce en application du règlement relatif aux statistiques européennes d'entreprises (règlement n° 2019/2152 du 27 novembre 2019[2]). Les estimations définitives de ces indices étaient – et sont toujours – calculées sur la base des seules données contenues dans les déclarations mensuelles de taxe sur la valeur ajoutée (TVA)[3] des entreprises, avec un décalage de deux mois par rapport à la période « sous revue » (période examinée). Ce décalage résulte d'une part du délai de réponse

2 Voir les références juridiques en fin d'article.
3 *https://www.insee.fr/fr/metadonnees/definition/c1777.*

▶ Encadré. Le schéma français de paiement par carte et par mobile CB : une spécificité française

Le schéma (ou réseau) domestique français de paiement par carte et mobile CB est organisé et piloté par le groupement d'intérêt économique Cartes Bancaires CB (ou GIE CB)*. Il a été créé en 1984 sous l'impulsion du ministère de l'Économie et des Finances, alors dirigé par Pierre Bérégovoy. Cette création accompagnait une forte dynamique d'innovation enclenchée par l'invention de la carte à puce, dix ans plus tôt (1974), et répondait à plusieurs objectifs :

• fédérer les deux blocs bancaires de l'époque – les banques mutualistes et les banques nationales – qui disposaient chacun de leur propre carte, pour doter la France d'un moyen de paiement interbancaire, électronique, sécurisé grâce à la carte à puce, et sous souveraineté française ;

• déployer sur l'ensemble du territoire des distributeurs automatiques de billets (DAB), ainsi que des terminaux de paiements électroniques chez les commerçants acceptant toutes les cartes bancaires CB, quelle que soit la banque émettrice.

En 40 ans, en nombre de transactions, la carte CB est devenue le moyen de paiement le plus utilisé en France par les consommateurs et les commerçants pour le paiement des dépenses de consommation courante, à la fois dans le commerce physique et dans le commerce en ligne. Elle est également le premier moyen d'accès aux espèces en France au travers du réseau de DAB bancaires CB.

En France, les cartes de paiement sont très majoritairement « cobadgées » : elles portent à la fois les marques du schéma français CB et d'un schéma de paiement international américain (Visa ou Mastercard). Le cobadgeage permet notamment aux porteurs français de pouvoir payer à l'étranger avec leur carte via les schémas Visa ou Mastercard. Il existe également des cartes émises en France ne portant que la marque d'un schéma international.

Les commerçants en France sont aussi très majoritairement affiliés à la fois au schéma CB et à différents schémas internationaux (Visa, Mastercard, Amex, etc.). Ceci leur permet notamment d'accepter les cartes de paiement qui ne comportent que la marque d'un schéma international.

Depuis l'entrée en vigueur, en 2016, du règlement européen n° 2015/751 relatif aux commissions d'interchange pour les opérations de paiement liées à une carte**, lorsque la carte est cobadgée et si le commerçant accepte les deux marques figurant sur la carte du consommateur, il appartient au commerçant et au consommateur de choisir la marque qu'ils souhaitent utiliser lors d'un paiement. Ceci est valable tant pour les paiements de proximité sur un terminal de paiement électronique que pour un paiement en ligne.

Le règlement européen permet au commerçant de présélectionner une des deux marques de la carte du porteur à condition que ce dernier puisse, s'il le souhaite, modifier ce choix lors du paiement (sauf impossibilité technique). En fonction du choix in fine réalisé, la transaction sera donc traitée soit par le schéma français CB, soit par le schéma international dont la marque figure également sur la carte.

En France, dans les transactions par carte ou par mobile entre un consommateur français et un commerçant français, compte tenu des choix opérés par les commerçants et les consommateurs, le schéma français de paiement par carte et par mobile CB reste le schéma de paiement le plus utilisé.

* *https://www.cartes-bancaires.com/cb/groupement/.*
** *Voir les références juridiques en fin d'article.*

accordé aux entreprises (un mois après la fin du mois sous revue) et d'autre part du temps de contrôle et d'apurement des données par l'Insee (un mois supplémentaire).

À l'époque, l'Insee disposait déjà des données de caisse[4] des enseignes de la grande distribution alimentaire, qu'il s'apprêtait à mobiliser de manière effective pour le calcul de l'indice des prix à la consommation (IPC)[5] (Leclair, 2019). Ces données sont exhaustives, mais sur un champ très restreint : celui des grandes surfaces principalement alimentaires. Aux fins du calcul précoce d'indices de chiffre d'affaires dans le commerce de détail, les données de paiement par carte constituent un complément d'information potentiellement intéressant, car elles couvrent l'ensemble du champ (y compris le commerce de détail non alimentaire et le commerce de détail alimentaire en petites surfaces). L'Insee a alors engagé des discussions avec le GIE CB. L'enjeu était que l'Insee puisse être destinataire d'agrégats de données de paiement par carte CB, susceptibles

4 *https://www.insee.fr/fr/metadonnees/definition/c2159.*
5 *https://www.insee.fr/fr/metadonnees/source/indicateur/p1653/description.*

> ***Avec l'accord des établissements bancaires membres du schéma CB et dans un cadre contractuel, l'Insee a pu accéder dès le printemps 2020 à des données préagrégées de flux de paiements CB.***

d'être mobilisés pour l'estimation avancée de l'indice de chiffre d'affaires du commerce de détail.

La crise sanitaire – avec la nécessité d'estimer rapidement l'ampleur du choc du confinement au printemps 2020 – a accéléré les discussions avec le GIE CB. Avec l'accord des établissements bancaires membres du schéma CB et dans un cadre contractuel, l'Insee a pu accéder dès le printemps 2020 à des données préagrégées de flux de paiements CB, sans possibilité de reconstituer les transactions individuelles, et donc sans risque de rupture de confidentialité vis-à-vis des porteurs de cartes et des commerçants. Les discussions se sont ensuite poursuivies et ont conduit en 2022 à l'entrée de l'Insee au sein de la chaire Finance digitale, hébergée à Télécom ParisTech et Paris 2 Panthéon-Assas. Dans le cadre de programmes de recherche conjoints CB-Insee, il est ainsi possible de mobiliser – dans des conditions extrêmement strictes de respect de la confidentialité et exclusivement à des fins de recherche – des données au niveau de détail le plus fin. L'accès aux données individuelles anonymisées, intermédié par les chercheurs habilités de la chaire, permet de mieux comprendre le mode de confection des données, leur portée et leurs limites, et d'élargir le champ des analyses effectuées avec ces données.

▶ Une couverture large, mais non exhaustive des transactions réalisées en France

L'unité statistique est la carte bancaire et la granularité la plus fine est la transaction réalisée par carte bancaire. Dans les données détaillées, on ne suit pas des individus mais uniquement des identifiants de carte bancaire, que seul le titulaire unique est autorisé contractuellement à utiliser, ce titulaire pouvant à l'inverse posséder plusieurs cartes.

Une part des cartes CB, certes minoritaire, est détenue par des entreprises ou des professionnels aux fins de régler des dépenses professionnelles. Or, les agrégats fournis par le GIE CB ne permettent pas de distinguer les transactions effectuées par des particuliers de celles qui sont réalisées par des entreprises.

Les paiements par carte, bien que largement utilisés, ne représentent qu'une partie des transactions monétaires. Les particuliers sont susceptibles d'utiliser les espèces, ce mode de paiement étant encore très fréquent dans le commerce physique[6], mais aussi les chèques, de manière plus marginale, ainsi que les virements bancaires ou les prélèvements, par exemple pour le règlement des factures d'électricité, d'eau, de télécommunications ou l'acquittement des loyers. Par ailleurs, certains paiements par carte réalisés en France ne sont pas traités par le schéma CB mais par des schémas internationaux (notamment Visa et Mastercard). C'est le cas en particulier

6 Comme évoqué précédemment, en 2024, 43 % des transactions réalisées dans des points de vente physiques l'ont été en espèces (Banque de France, 2025).

> ***Ces données permettent de compléter les statistiques habituellement produites par l'Insee, en apportant des informations plus précoces et plus fines, ou en permettant d'éclairer certains angles d'analyse inabordables avec les données classiques.***

des transactions réalisées en France par des clients étrangers ou par des porteurs de cartes non CB et, de manière symétrique, par des porteurs français à l'étranger. C'est également le cas des transactions domestiques avec des cartes cobadgées lorsque le choix opéré par le commerçant et/ou le porteur est en faveur d'un schéma international.

Ainsi, les transactions enregistrées par CB ne le sont qu'à hauteur de la part de marché de CB parmi les différents schémas de paiement par carte présents en France. Par rapprochement avec diverses sources, on estime que CB représente environ 70 % des montants payés par carte en France en 2023[7]. Ce taux est cependant susceptible d'évoluer à la baisse ou à la hausse en fonction des dynamiques concurrentielles entre les schémas (depuis 2021, la part de marché de CB a diminué au sein des schémas de paiement en France, mais l'évolution observée au cours des dernières années ne préjuge pas des évolutions futures).

En dépit de leur champ non exhaustif, ces données permettent de compléter les statistiques habituellement produites par l'Insee, en apportant des informations plus précoces et plus fines, ou en permettant d'éclairer certains angles d'analyse inabordables avec les données classiques.

▶ Des traitements nécessaires pour permettre un usage à des fins de statistique publique

Les données de paiement par carte ne sont pas structurées pour les besoins de la statistique publique, mais selon des besoins de pilotage du schéma CB, de traçabilité, de lutte contre la fraude, ainsi que des besoins d'information pour les relevés de compte des commerçants et des porteurs de carte.

Les données de paiement CB mobilisées par le GIE CB dans le cadre de la coopération contractualisée entre l'Insee et le GIE CB proviennent de plusieurs sources :

- Les **données d'autorisation pour les paiements CB**, qui remontent en temps réel au GIE CB. Elles constituent un **échantillon** des transactions effectives. En effet, dans le commerce physique, les paiements CB peuvent être réalisés « *offline* », c'est-à-dire sans demande d'autorisation. Les modalités de déclenchement de l'autorisation lors d'une transaction de paiement de proximité suivent des règles de sécurité définies à la fois par le GIE CB, l'établissement bancaire du commerçant et l'établissement bancaire émetteur de la carte. De plus, dans certains cas d'usage, le montant de l'autorisation peut être différent du montant final de la transaction financière, par exemple dans le cas des distributeurs automatiques de carburants ou encore dans celui des paiements

7 Ce chiffre a été calculé avec les données mensuelles d'activité de paiements CB (voir infra) et un tableau de l'Observatoire de la sécurité des moyens de paiement (2024).

pour la location de biens ou de services (empreinte bancaire). Au total, cet échantillon offre une couverture étendue et sa qualité permet une analyse pertinente, même si sa représentativité peut présenter certaines limites mineures (en particulier au niveau sectoriel).

- Les **données de compensation pour les paiements CB**, qui remontent aussi en temps réel au GIE CB. Elles constituent également un **échantillon** des transactions effectives. En effet, il n'y a compensation que lorsque la banque du commerçant diffère de celle de l'acheteur porteur de carte : on parle alors de transaction interbancaire. Lorsque ces deux banques ne font qu'une, celle-ci enregistre la transaction dans son système d'information, en créditant le compte du commerçant et en débitant celui de l'acheteur, sans que la transaction soit échangée dans le système de compensation : on parle alors de transaction intrabancaire. Au total, cet échantillon offre une couverture étendue et sa qualité permet une analyse pertinente, même si sa représentativité peut présenter certaines limites mineures (en particulier au niveau local, si une banque a une situation monopolistique).

- Les **données d'activité de paiements CB agrégées sur une base mensuelle**, ventilées par commerçant affilié à CB. Elles incluent les transactions de paiement CB autorisées ou non autorisées, interbancaires et intrabancaires. Cette source rassemble toutes les transactions effectives, elle est donc **exhaustive**.

Les traitements réalisés par le GIE CB en amont de la transmission à l'Insee

Pour être utilisables à des fins de statistique publique, les données détaillées recueillies par le GIE CB doivent subir différents traitements : filtrage, enrichissement par appariement avec d'autres fichiers, retraitement éventuel et, enfin, agrégation. En particulier, l'enrichissement vise à compléter l'information sur les commerçants concernés par les transactions.

> ***Pour être utilisables à des fins de statistique publique, les données détaillées recueillies par le GIE CB doivent subir différents traitements : filtrage, enrichissement par appariement avec d'autres fichiers, retraitement éventuel et, enfin, agrégation.***

Initialement, les données d'une transaction de paiement CB comprennent différentes données sur le commerçant concerné émanant de son établissement bancaire : son identification (numéro Siret), sa localisation géographique et son code d'activité selon une nomenclature spécifique à la monétique : le *Merchant Category Code* (MCC)[8]. Avant d'être communiquées à l'Insee, les données d'autorisation et de compensation, ainsi que celles de l'activité mensuelle des paiements, sont enrichies par le GIE CB du code de l'activité principale exercée (APE)[9] selon la nomenclature d'activités française (NAF)[10].

8 Nomenclature internationale utilisée par les acteurs de la monétique.
9 *https://www.insee.fr/fr/metadonnees/definition/c1888.*
10 *https://www.insee.fr/fr/information/2406147.*

Le code APE est obtenu par appariement avec le répertoire Sirene[11] tenu par l'Insee. La qualité de l'appariement est tributaire du rattachement de la transaction à la bonne entreprise dans le système d'information de la banque. Notamment, l'appariement est problématique si un groupe qui détient plusieurs magasins associe tous ses points de vente à son siège social alors que ce dernier n'est pas classé dans le commerce.

Les données agrégées transmises régulièrement à l'Insee par le GIE CB

Le GIE CB transmet précocement et périodiquement à l'Insee les **agrégats journaliers d'autorisation et de compensation** pour les paiements CB (nombre de transactions et montant total par jour). Des mises à jour successives sont effectuées pour intégrer les ajustements et corrections éventuels du réseau de paiement. Ces corrections, généralement modérées, représentent en moyenne moins de 1 % du montant total quotidien. Les deux types d'agrégats sont ventilés selon les trois variables suivantes : le mode de paiement (paiement dans le commerce physique en mode contact ou sans contact, paiement en ligne), le département du commerçant et son code d'activité principale (MCC ou APE). Pour chacun des deux types d'agrégats, le GIE CB transmet un fichier par variable de ventilation. Il communique en outre tous les mois le **cumul mensuel des données d'activité des paiements CB**, ventilé selon le croisement entre le département du commerçant et le code MCC ou APE de son établissement, offrant ainsi une vision consolidée des tendances de paiement.

L'historique des données fournies à l'Insee remonte à janvier 2018 pour l'activité mensuelle des paiements CB et à janvier 2019 pour les agrégats journaliers des données d'autorisation et de compensation, offrant ainsi un aperçu sur plusieurs années des paiements par carte bancaire CB.

Travailler sur des données préagrégées nécessite un échange régulier avec le producteur

Le statisticien ne procède pas lui-même aux traitements précités. Pour bien saisir les limites des données agrégées ainsi constituées, il se doit d'acquérir auprès du producteur (le GIE CB) une bonne connaissance de la méthodologie mise en œuvre par celui-ci. La qualité de la relation et des échanges entre les statisticiens publics et les experts de la donnée du côté du producteur est donc essentielle.

L'accès à des données agrégées plutôt qu'à des données détaillées ne présente pas que des inconvénients pour le statisticien. En effet, les données détaillées sont d'une volumétrie très importante, et ne sont ni structurées, ni documentées pour les usages de la statistique publique.

11 *https://www.insee.fr/fr/metadonnees/source/serie/s1020.*

▶ Les contrôles de qualité réalisés à l'Insee sur les agrégats transmis par le GIE CB

Le transfert des données s'effectue de manière sécurisée via la plateforme d'échange de fichiers de l'Insee, garantissant l'intégrité et la confidentialité des informations.

Les fichiers transmis à l'Insee font l'objet de plusieurs contrôles pour garantir leur conformité, leur qualité et leur fiabilité. En plus des contrôles classiques (plage temporelle complète et valide, présence et format des variables, stabilité des modalités, etc.), l'institut s'assure de la cohérence entre les agrégats d'autorisation, de compensation et d'activité mensuelle (par comparaison des totaux, des répartitions et des glissements temporels). Enfin, des contrôles de vraisemblance sont appliqués : ils visent à détecter des valeurs atypiques qui pourraient signaler des erreurs grossières ou des biais importants dans les données. Il s'agit d'abord de contrôles de vraisemblance interne pour repérer les variations significatives des glissements temporels ; ces variations doivent être interprétées en tenant compte de la saisonnalité des séries. Des contrôles de vraisemblance externe sont également réalisés en mobilisant la source TVA : cette source est mensuelle, disponible relativement rapidement, quoiqu'un peu plus tardivement que la source CB, et se rapproche de cette dernière tant du point de vue des concepts (approche sectorielle notamment) que des variables (activité ou consommation sur tel ou tel champ). Les données de caisse sont également mobilisées, mais pour le champ restreint du commerce en grande surface. Tous ces contrôles sont réalisés chaque semaine après une nouvelle livraison et donnent lieu à un rapport automatisé, permettant ainsi un suivi constant de la qualité des données.

> ***Les fichiers transmis à l'Insee font l'objet de plusieurs contrôles pour garantir leur conformité, leur qualité et leur fiabilité.***

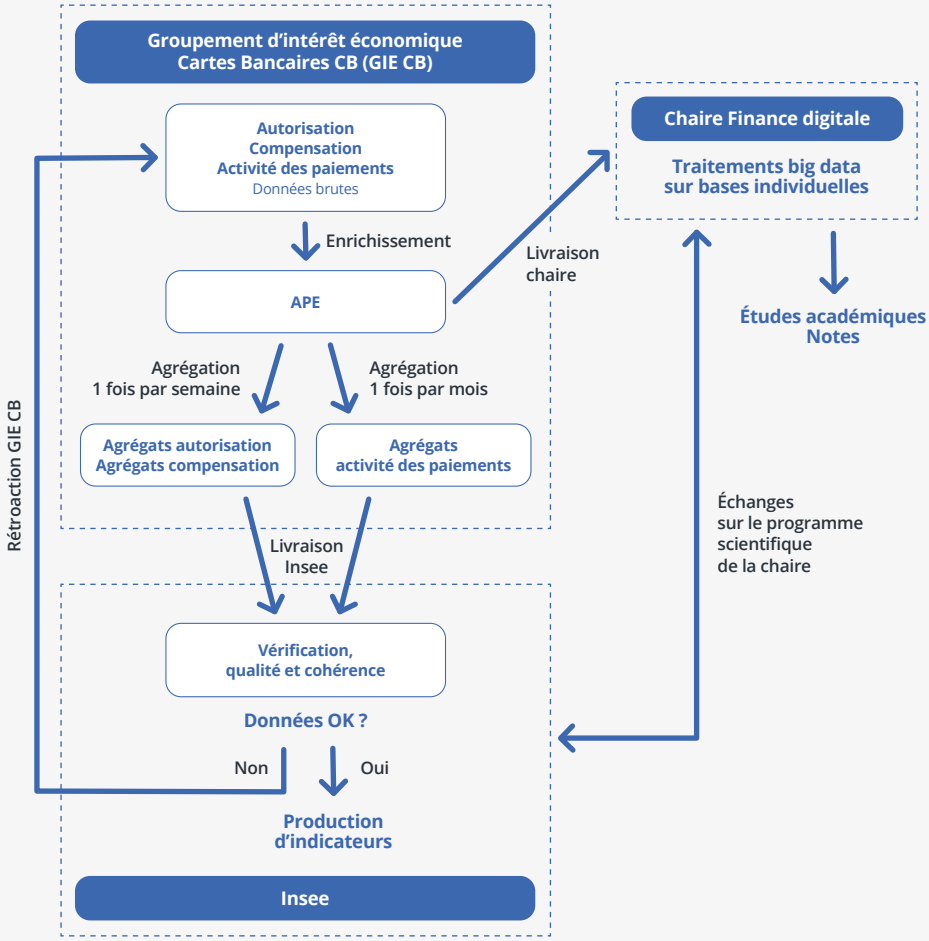
L'Insee ne réalise aucun retraitement de ces données. En cas d'anomalie détectée lors des contrôles, une demande d'expertise, de correction et de relivraison est effectuée auprès du GIE CB. Une fois les nouvelles données validées, elles sont mises en forme et concaténées avec l'historique, garantissant une structure homogène et facilitant l'exploitation. Ce fichier final est ensuite mis à disposition des utilisateurs internes à l'Insee (***figure 4***).

▶ Une source d'information complémentaire sur la consommation, l'activité commerciale…

Un apport limité pour le suivi conjoncturel et la prévision de la consommation en dehors de la période de la crise sanitaire

La mise à disposition rapide des données CB et leur caractère inframensuel en font un outil particulièrement intéressant pour le suivi conjoncturel et le *nowcasting*, c'est-à-dire l'estimation en temps réel des comportements des agents économiques. Elles peuvent notamment être mobilisées pour les prévisions de consommation des

Figure 4 - Représentation des flux de données : du GIE CB au statisticien



APE : Activité principale exercée.

ménages présentées dans la *Note de conjoncture* de l'Insee. Cette dernière est publiée en général au milieu du troisième mois d'un trimestre, les prévisions démarrant ce même trimestre. Or, à la fin d'un mois donné, les informations sur les transactions ayant eu lieu au cours des trois premières semaines via le schéma CB sont disponibles – bien que les informations portant sur les deuxième et troisième semaines puissent être révisées de façon mineure (voir supra). Ainsi, les données CB peuvent permettre d'estimer les deux premiers mois du trimestre en cours.

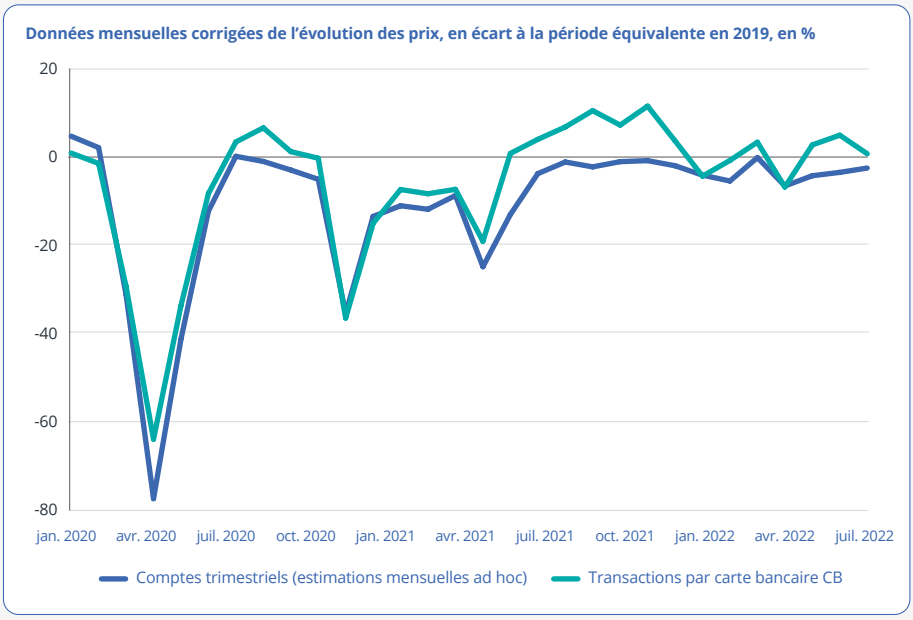
La totalité de la consommation des ménages ne passe cependant pas par des transactions par carte bancaire CB (voir supra). Pendant la crise sanitaire, les postes pour lesquels l'évolution des montants des transactions par carte bancaire CB s'est avérée être un bon

> ***La mise à disposition rapide des données CB et leur caractère inframensuel en font un outil particulièrement intéressant pour le suivi conjoncturel et le* nowcasting*, c'est-à-dire l'estimation en temps réel des comportements des agents économiques.***

prédicteur de la consommation des ménages, telle que retracée par les comptes nationaux trimestriels, représentaient environ 18 % du total de la consommation des ménages. En particulier, les mouvements de la consommation des ménages dans la restauration (secteur qui représente 6 % des dépenses des ménages au total) ou les carburants (3 % des dépenses) ont été correctement représentés par la dynamique des transactions par carte bancaire CB, dans un contexte où les évolutions au mois le mois atteignaient plusieurs dizaines de points de pourcentage (***figure 5***).

En revanche, dans une période où les évolutions de la consommation des ménages ne sont, à l'échelle d'un trimestre, que de l'ordre de quelques dixièmes de points de pourcentage, les données de transactions par carte bancaire CB n'améliorent pas significativement la prévision par rapport aux outils traditionnels qui mobilisent les enquêtes de conjoncture (qualitatives) auprès des ménages et des entreprises. En effet, le « bruit »[12] contenu dans les données CB l'emporte alors sur le « signal » qu'elles pourraient apporter.

▶ Figure 5 - Consommation des ménages en carburants, en comptabilité nationale et dans les transactions par carte bancaire CB



Sources : Cartes Bancaires CB, activité des paiements CB ; Insee, comptes trimestriels ; calculs Insee.

12 Fluctuations non directement liées au phénomène que l'on cherche à analyser.

L'estimation avancée du volume des ventes dans le commerce de détail

La production par l'Insee d'indicateurs mensuels de suivi de l'activité répond à des exigences européennes (règlement relatif aux statistiques européennes d'entreprises[13]), à des besoins nationaux (suivi conjoncturel, établissement des comptes trimestriels et, notamment, première estimation du produit intérieur brut (PIB)[14]), et à la nécessité de diffuser une information économique conjoncturelle de qualité aux agents économiques qui en ont besoin.

L'Insee, en réponse aux exigences européennes, doit transmettre 30 jours après la fin du mois sous revue un indice provisoire du volume des ventes du commerce de détail. L'indice définitif doit être transmis à +60 jours. Pour l'estimation à +30 jours, en l'absence de source concurrente sur un champ comparable, les données CB sont disponibles suffisamment précocement et offrent un proxy acceptable des ventes du commerce de détail ; les données administratives ne sont pas accessibles dans un tel délai. On calcule des sous-indices à un niveau de détail plus fin que celui requis, puis on les agrège avec les pondérations issues de l'indice définitif à +60 jours ; ceci permet de corriger des écarts de couverture et d'améliorer la fiabilité des estimations provisoires. La qualité des estimations ainsi obtenues a été jugée suffisamment bonne par rapport à l'ancienne méthode, qui mobilisait les données de l'enquête mensuelle sur l'activité des grandes surfaces alimentaires (*Emagsa*[15]). Aussi, il a été décidé d'abandonner cette enquête et de réduire ainsi la charge statistique des entreprises. En outre, les données CB, conjuguées à d'autres sources, permettent d'envisager, à terme, d'affiner la granularité des estimations précoces, voire d'étendre leur usage à d'autres secteurs d'activité où le paiement par carte est fréquent.

▶ … ou les dépenses touristiques à une maille géographique fine

Les données de compensation et les données d'activité mensuelle des paiements CB peuvent être utilisées pour les analyses conjoncturelles aux niveaux régional et départemental, en complément d'autres données (comme les données expérimentales d'indices régionaux de chiffres d'affaires ou les estimations trimestrielles d'emploi localisées). Les données de compensation permettent de suivre le total des transactions départementales dans les délais les plus courts, mais les données d'activité mensuelle permettent en plus une approche sectorielle, notamment sur le tourisme. Les données d'activité mensuelle comprennent par ailleurs l'ensemble des transactions par carte bancaire CB (y compris en intrabancaire) : elles offrent donc un gain de précision.

Dès 2020, les données de compensation ont permis de suivre, au niveau des départements, les fortes évolutions des montants des transactions par carte bancaire CB pendant les périodes de confinement associées à la crise sanitaire, éclairant ainsi les disparités territoriales de consommation (Insee, 2020). Entre la semaine du 2 au 8 mars 2020 et celle du 23 au 29 mars 2020, le montant global des transactions a chuté de

13 Voir les références juridiques en fin d'article.
14 *https://www.insee.fr/fr/metadonnees/definition/c1365.*
15 *https://www.insee.fr/fr/metadonnees/source/serie/s1222.*

façon très marquée dans les départements de Paris, de la Savoie, des Hautes-Alpes et des Hautes-Pyrénées. Ces évolutions ont été expliquées par les mouvements de population observés à l'annonce du confinement et par la chute de la fréquentation touristique, liée entre autres aux fermetures anticipées des stations de sports d'hiver.

> ***Les disparités territoriales étant plus importantes dans le secteur du tourisme, le principal intérêt des données d'activité mensuelle des paiements CB est de pouvoir suivre les transactions dans l'hébergement et la restauration au niveau départemental.***

Les disparités territoriales étant plus importantes dans le secteur du tourisme, le principal intérêt des données d'activité mensuelle des paiements CB est de pouvoir suivre les transactions dans l'hébergement et la restauration au niveau départemental. Par exemple, au cours des étés 2020 et 2021, on a observé une remontée plus importante des paiements par carte bancaire CB que des chiffres d'affaires issus de la source fiscale à Paris et dans les Alpes-Maritimes, ce qui suggère que ces deux départements ont été particulièrement concernés par les limitations imposées à la venue de touristes étrangers (***figure 6***). Les restrictions de déplacements internationaux mises en place du fait de la crise sanitaire avaient alors conduit davantage de résidents à passer leurs vacances en France, tandis que les touristes internationaux n'étaient pas encore revenus. Les fortes dépenses des résidents n'ont cependant pas compensé l'absence des touristes étrangers.

Toutefois, hors périodes de fortes évolutions, les données de transactions par carte bancaire CB n'apportent pas de complément d'information substantiel par rapport aux autres sources conjoncturelles régionales. De fait, les transactions dans l'hébergement couvrent un champ moins étendu que celui des données de chiffres d'affaires, notamment parce que les transactions effectuées à l'avance auprès des plateformes de réservation sur Internet ne sont pas attribuées à ce secteur. En outre, comme l'ensemble des transactions en ligne ne peuvent pas être correctement localisées (voir infra), elles sont exclues des données départementales.

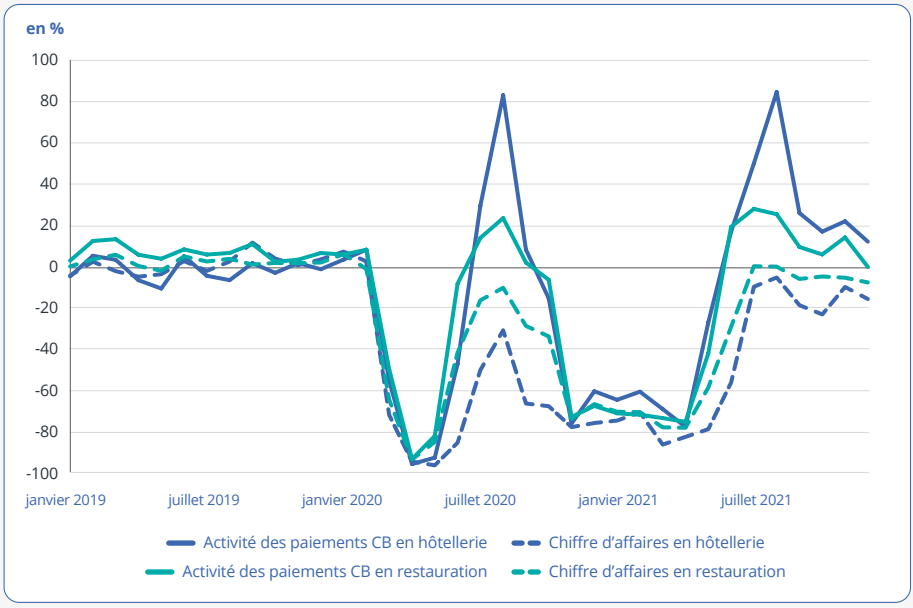
▶ Les déplacements des porteurs de carte constituent un bon indicateur des connexions entre commerces et entre territoires

Les données individuelles de transaction par carte bancaire CB (mobilisables de façon sécurisée et anonymisée exclusivement dans le cadre du programme de recherche CB-Insee de la chaire Finance digitale) permettent de suivre les transactions réalisées par une même carte. On peut ainsi établir des chaînes de transactions successives dans des commerces de détail distincts. Ce chaînage permet d'étudier des questions de recherche comme :

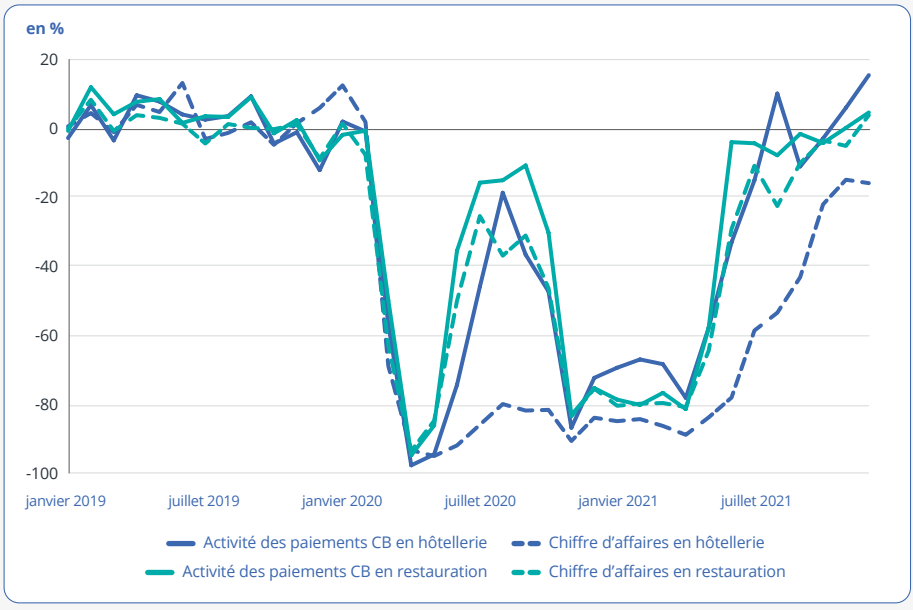
- Quels sont les commerces les plus connectés entre eux et où sont-ils localisés ?

▶ Figure 6 - Évolution des montants d'activité des paiements CB et de chiffre d'affaires en 2020 et 2021, par rapport au même mois de 2019, et en 2019 par rapport au même mois de 2018

a Dans le département des Alpes-Maritimes



b Dans le département de Paris



Sources : Cartes Bancaires CB, montants d'activité des paiements CB ; DGFiP, déclarations de chiffre d'affaires (TVA) ; calculs Insee.

• Comment la présence d'un grand centre commercial en périphérie influe-t-elle sur les comportements de consommation dans les petits commerces de centre-ville ?

• Quels sont les commerces ou secteurs qui attirent le plus de consommateurs issus de communes éloignées ?

L'éclairage apporté par cette nouvelle source dans le cadre de ces travaux de recherche académique complète les analyses effectuées à partir des sources traditionnelles de la statistique publique. En effet, les études déjà existantes caractérisent les pôles commerçants à partir de la densité des commerces, indépendamment de leur fréquentation effective par la population. Elles s'intéressent à l'évolution des effectifs salariés des points de vente dans ces pôles, relativement au reste de la commune (Cazaubiel et Guymarc, 2019), en lien éventuellement avec les caractéristiques de l'aire urbaine (Bessière et Trevien, 2016).

Les chaînes de transactions permettent aussi d'établir des liens entre territoires. Par exemple, si beaucoup de porteurs de carte font leurs achats successivement dans une commune, puis une autre, on peut supposer qu'il existe une forte connexion entre ces deux communes.

Les chaînes de transactions permettent aussi d'établir des liens entre territoires. Par exemple, si beaucoup de porteurs de carte font leurs achats successivement dans une commune, puis une autre, on peut supposer qu'il existe une forte connexion entre ces deux communes. Les connexions entre territoires sont actuellement considérées à partir des déplacements domicile-travail (aires d'attraction des villes[16]), ou en fonction des déplacements nécessaires pour accéder aux équipements et services courants les plus proches (bassins de vie[17]), indépendamment des comportements effectifs de la population. Les déplacements liés à la consommation présentent l'intérêt de concerner une large fraction de la population, et pas uniquement les actifs occupés, d'élargir ainsi les motifs de déplacements et de donner une précieuse indication sur la fréquentation des équipements.

Au-delà de ces premiers usages réalisés ou envisagés en France, dans d'autres pays, plusieurs organismes institutionnels et académiques ont mis en place des collaborations avec les acteurs de référence du secteur bancaire afin d'utiliser les données de transactions par carte. Par exemple, au Royaume-Uni, l'autorité de la concurrence (*Competition and Markets Authority* – CMA) et l'institut de statistique (*Office for National Statistics* – ONS) utilisent les données Visa pour étudier la géographie du marché de commerce de détail (Doshi et al., 2024). Ou encore, en Allemagne, Alipour et al. (2022) estiment l'impact de la montée en puissance du télétravail sur la consommation dans 50 villes, grâce à une combinaison des données de téléphonie mobile et des données de transaction par carte bancaire fournies par Mastercard.

16 *https://www.insee.fr/fr/metadonnees/definition/c2173.*
17 *https://www.insee.fr/fr/metadonnees/definition/c2060.*

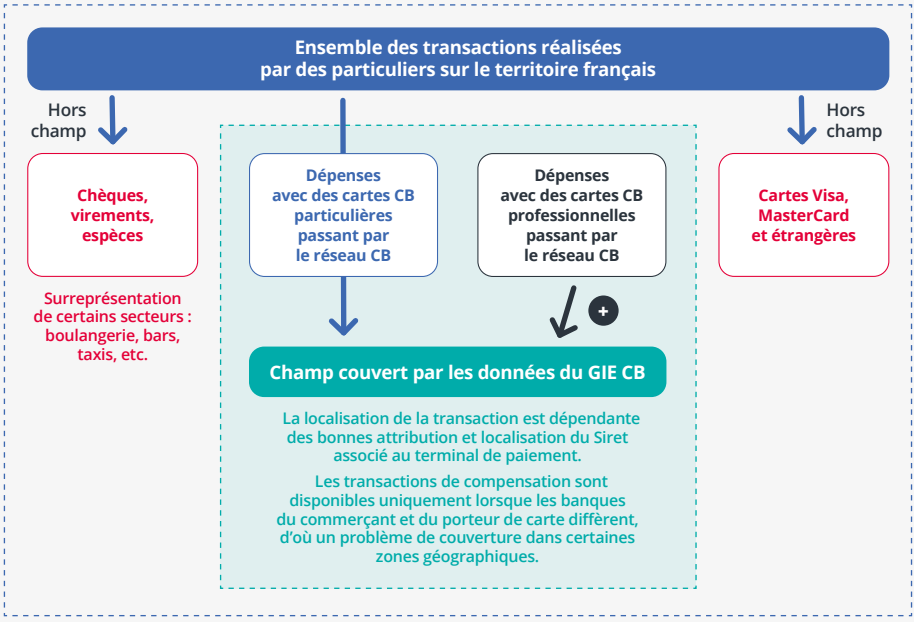
▶ Des limites intrinsèques au champ des données et à leur contenu

L'utilisation des données de paiement par carte CB à des fins de statistiques publiques ou d'études se heurte à deux types de limites : d'une part, celles inhérentes au champ des données, à leur objet et à leur mode de confection, d'autre part, celles résultant des comportements des porteurs de cartes, des commerçants, des banques et des schémas de paiement.

Pour ce qui est des limites inhérentes au champ des données (***figure 7***), on rappelle d'abord qu'elles ne couvrent que des paiements réalisés dans le cadre du schéma CB, c'est-à-dire entre un commerçant affilié au schéma CB et un porteur d'une carte CB, le commerçant et/ou le porteur ayant par ailleurs choisi la marque CB dans le cas où la carte est cobadgée. Globalement, cela correspond aux paiements effectués par des personnes résidant en France, qui disposent d'une carte CB ou cobadgée CB, et qui réalisent un achat dans un point de vente en France ou sur un site e-commerce s'adressant aux consommateurs en France.

Par ailleurs, si l'on cherche à mesurer les évolutions de la consommation des ménages, l'évolution des paiements par carte CB n'est pas parfaitement adaptée. En effet, comme on l'a vu, une part desdits paiements, certes très minoritaire, correspond à des dépenses prises en charge par des entreprises.

▶ Figure 7 - Ensemble des transactions des particuliers et part couverte par le groupement d'intérêt économique Cartes Bancaires CB (GIE CB)



De plus, les données CB ne couvrent pas les paiements en espèces ou par chèque. Or, l'usage du liquide ou du chèque est encore fréquent dans certains secteurs d'activités (boulangeries, bars, taxis, etc.), de sorte que le taux de couverture des paiements par carte CB varie d'un secteur à un autre.

On rappelle également que les données d'autorisation ne correspondent pas toujours à des transactions effectives, et que les données de compensation ne couvrent pas les transactions entre un particulier et un commerçant qui sont clients de la même banque.

Enfin, le niveau de couverture peut être affecté par des changements réglementaires ou « métier » liés au paiement. Cela a été le cas, par exemple, en 2022, lorsque le seuil de paiement sans contact a été relevé de 30 euros à 50 euros. Ce changement a eu pour effet de permettre – et donc d'encourager – le paiement par carte pour des achats compris entre ces deux montants.

▶ L'activité et la localisation sont connues de manière imparfaite

Le code d'activité principale du commerçant, qu'il s'agisse du code APE ou du code MCC, n'est pas essentiel au dénouement de la transaction, si bien qu'il est parfois erroné. Le code APE correspond souvent à celui qui avait été attribué par l'Insee à l'entreprise au démarrage de son activité, car il est mis à jour ensuite de manière irrégulière[18]. Quant au code MCC, il arrive qu'un commerçant fasse usage d'un code unique pour des activités différentes. Le cas de figure est fréquent, par exemple, pour les stations-service exploitées par un supermarché ou un hypermarché : l'ensemble des recettes est parfois, mais pas toujours, regroupé sous le code 5411 « Épiceries, supermarchés », sans qu'il soit possible d'isoler les recettes en carburants.

Le code d'activité principale du commerçant, qu'il s'agisse du code APE ou du code MCC, n'est pas essentiel au dénouement de la transaction, si bien qu'il est parfois erroné.

Pour des utilisations des données CB au niveau régional ou départemental, la localisation attribuée au commerçant peut, elle aussi, être erronée. Les ventes réalisées par une enseigne commerciale ayant une implantation nationale sont dans certains cas toutes localisées au siège de l'enseigne.

▶ Les limites inhérentes aux évolutions comportementales

Le taux de couverture des dépenses par les données CB est également susceptible d'être affecté par des changements de comportement des banques, des commerçants et des porteurs de carte :

18 Ce peut être par exemple à la suite d'une demande de révision par l'entreprise, selon la démarche explicitée sur le site de l'Insee (*https://www.insee.fr/fr/information/7614104*).

> ***Le taux de couverture des dépenses par les données CB est également susceptible d'être affecté par des changements de comportement des banques, des commerçants et des porteurs de carte.***

• Les banques émettrices peuvent décider de ne pas ou plus faire figurer la marque CB sur tout ou partie de leurs cartes et de ne faire figurer que celle d'un schéma international (on parle alors de « décobadgeage »). Elles peuvent aussi, à l'inverse, décider de cobadger des cartes qui ne l'étaient pas auparavant.

• Les commerçants peuvent choisir de ne plus accepter certains schémas de paiement par carte ou de modifier, pour une période donnée, le schéma privilégié, que ce soit en faveur de CB ou, à l'inverse, en faveur d'une marque internationale. Des nouvelles enseignes étrangères peuvent s'installer en France, n'accepter dans un premier temps que les schémas internationaux, puis décider d'accepter aussi le schéma CB.

• Les porteurs de carte peuvent faire un usage plus ou moins large de leur carte compte tenu des autres modes de paiement possibles. Ils peuvent par ailleurs opter pour des cartes cobadgées ou des cartes non cobadgées CB en fonction de ce que leur propose leur banque. Les porteurs de cartes cobadgées peuvent aussi, s'ils le souhaitent, modifier la marque présélectionnée par le commerçant et choisir l'autre marque pour les paiements de proximité (sauf impossibilité technique) ou pour un paiement sur internet.

Ces dynamiques concurrentielles conduisent à des fluctuations de la part de marché de CB, et donc de la couverture de cette source. Ces dernières peuvent être significatives, à la baisse ou à la hausse, ce qui affecte in fine les interprétations possibles des évolutions des phénomènes que l'on souhaite appréhender par les données.

Afin de mesurer la couverture des données CB, la source TVA évoquée précédemment est mobilisée en comparaison. Il s'agit de la source principale utilisée par l'Insee pour calculer les indices mensuels définitifs de chiffre d'affaires par activité. À partir de cette source, on estime un chiffre d'affaires mensuel total par secteur d'activité, auquel on rapporte le montant mensuel total des transactions par carte CB. Ce ratio représente une estimation de la couverture de la source CB parmi l'ensemble des moyens de paiement (carte bancaire ou autre) et des réseaux de paiement (CB ou autre). L'analyse de ce taux de couverture permet dans un premier temps d'identifier les secteurs bien couverts dans le commerce de détail, qui sont ceux pour lesquels il peut être pertinent d'utiliser la source CB, puis dans un second temps d'appréhender les distorsions dues aux changements comportementaux ou réglementaires.

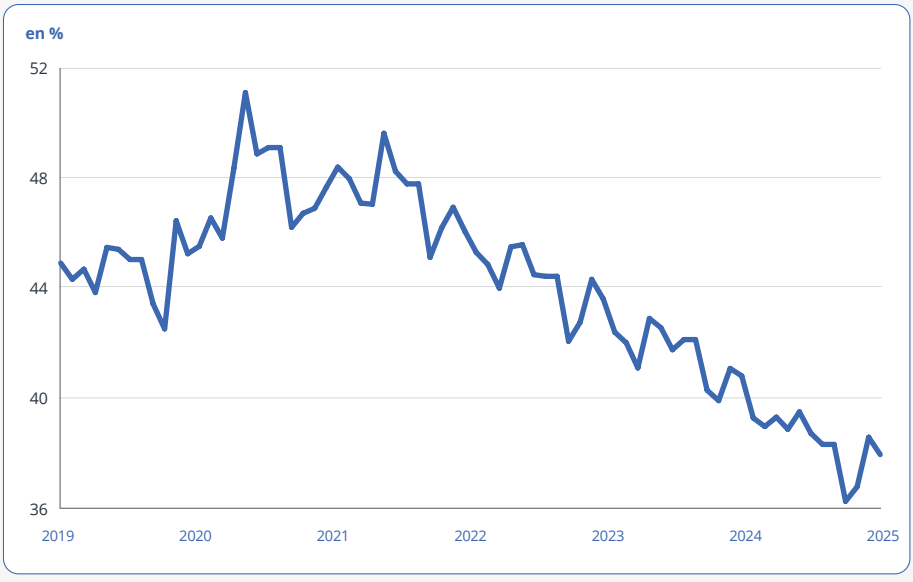
En 2023, les paiements par carte CB ont représenté environ 42 % des dépenses totales dans le commerce de détail en France. Cependant, on constate des variations significatives du taux de couverture, qui peuvent être dues à trois types de facteurs : la saisonnalité, la volatilité de court terme et les changements structurels (qu'il s'agisse de biais tendanciels ou de ruptures) :

• La **saisonnalité** est en lien avec la part des dépenses par carte parmi les moyens de paiement, mais sans lien avec la part du réseau de paiement CB par rapport à ses concurrents. Par exemple, les périodes de soldes, comme celles d'hiver et d'été, engendrent des hausses des dépenses pour des produits plus souvent achetés par carte CB, ce qui peut induire une augmentation de la couverture.

• La **volatilité** de court terme peut avoir plusieurs origines : variabilité de l'échantillonnage (pour les données d'autorisation), problèmes de mesure, fluctuations dans la production des données du côté du GIE CB, différences de champ avec la source de référence, etc. Elle est difficilement quantifiable.

• Les **changements structurels** représentent quant à eux les dynamiques de long terme. Ainsi, on observe une augmentation du taux de couverture sur la période 2020-2021, liée notamment à des changements de comportement engendrés par la crise sanitaire (par exemple la baisse de l'usage des espèces en faveur du paiement par carte) et au relèvement du plafond de paiement sans contact, en particulier pendant et après le premier confinement, de mars à mai 2020 (***figure 8***). Depuis 2021, la couverture CB a subi en revanche une baisse, causée en partie par des facteurs concurrentiels : la part de marché de CB a diminué au sein des schémas de paiement en France. Ce biais d'évolution, difficilement contrôlable sauf en cas d'évolution tendancielle régulière, suggère qu'il est plus pertinent d'utiliser les données CB (désaisonnalisées) sur une période courte, et pour des chocs de grande ampleur, plutôt que pour appréhender des évolutions de moyen terme.

▶ Figure 8 - Taux de couverture des paiements par carte bancaire CB sur le champ du commerce de détail (NACE G47)



Sources : Cartes Bancaires CB, données de compensation pour les paiements par carte CB ; DGFiP, déclarations de chiffre d'affaires (TVA) ; calculs Insee.

▶ Perspectives

Au total, les données de paiement par carte CB constituent une source mobilisable – parmi d'autres – pour construire une estimation précoce d'indices de consommation et d'indices d'activité dans le commerce de détail, ou pour appréhender des chocs de grande ampleur en la matière. Elles ne constitueront sans doute pas la source privilégiée – a fortiori la source unique – pour l'élaboration de ces indicateurs, ne serait-ce qu'en raison de leur champ partiel et des fluctuations du taux de couverture imputables aux évolutions des comportements des ménages, des banques, des commerçants et des schémas de paiement. Si ces inconvénients sont rédhibitoires pour l'élaboration des valeurs définitives des indices, la disponibilité rapide des données CB les rend néanmoins très utiles pour des estimations précoces.

Les données CB constituent aussi une source très prometteuse pour éclairer des problématiques de recherche relatives aux comportements de consommation, de localisation ou de déplacements des ménages. Pour ce faire, les données peuvent être utilisées isolément ou, mieux encore, combinées à d'autres sources, comme les données de téléphonie mobile[19].

19 Voir l'article de Marie-Pierre Joubert sur les données de téléphonie mobile dans ce même numéro.

▶ Fondements juridiques

- Règlement (UE) n° 2015/751 du Parlement européen et du Conseil du 29 avril 2015 relatif aux commissions d'interchange pour les opérations de paiement liées à une carte. In : *site de l'Union européenne*. [en ligne]. [Consulté le 29 janvier 2025]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=celex:32015R0751*.

- Règlement (UE) n° 2019/2152 du Parlement européen et du Conseil du 27 novembre 2019 relatif aux statistiques européennes d'entreprises, abrogeant dix actes juridiques dans le domaine des statistiques d'entreprises. In : *site de l'Union européenne*. [en ligne]. [Consulté le 29 janvier 2025]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/?qid=1730207160102&uri=CELEX%3A32019R215*.

▶ Bibliographie

- ALIPOUR, Jean-Victor, FALCK, Oliver, KRAUSE, Simon, KROLAGE, Carla et WICHERT, Sebastian, 2022. Working from Home and Consumption in Cities. In : *CESifo Working Paper*. [en ligne]. Octobre 2022. N° 1000. [Consulté le 29 janvier 2025]. Disponible à l'adresse : *http://dx.doi.org/10.2139/ssrn.4255438*.

- BANQUE DE FRANCE, 2025. Les Français continuent d'apprécier les espèces, même si leur usage se réduit au profit des paiements par carte et mobile. In : *site de la Banque de France*. [en ligne]. 25 février 2025. [Consulté le 29 janvier 2025]. Disponible à l'adresse : *https://www.banque-france.fr/fr/publications-et-statistiques/publications/les-francais-continuent-dapprecier-les-especes-meme-si-leur-usage-se-reduit-au-profit-des-paiements*.

- BESSIÈRE, Sabine et TREVIEN, Corentin, 2016. Le commerce de centre-ville : une vitalité souvent limitée aux grandes villes et aux zones touristiques. In : Insee *Références*. [en ligne]. 8 novembre 2016. Insee. [Consulté le 29 janvier 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/2497068?sommaire=2497179*.

- CAZAUBIEL, Arthur et GUYMARC, Gaël, 2019. La déprise du commerce de proximité dans les centres-villes des villes de taille intermédiaire. In : *Insee Première*. [en ligne]. 14 novembre 2019. Insee. N° 1782. [Consulté le 29 janvier 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4248184*.

- CAZAUBIEL, Arthur, DARMAILLACQ, Corinne, LEBLANC, Pierre, CHEPTITSKI, Aliette et SIMON, Olivier, 2022. Apports, limites et perspectives des données de transactions carte bancaire (CB) dans le suivi de l'activité économique. In : *site des JMS*. [en ligne]. [Consulté le 29 janvier 2025]. Disponible à l'adresse : *https://journees-methodologie-statistique.insee.net/apports-limites-et-perspectives-des-donnees-de-transactions-carte-bancaire-cb-dans-le-suivi-de-lactivite-economique/*.

- DOSHI, Samir, HOOLOHAN, Vicky, LEWIS, Tabitha et SCHNEEBACHER, Jakob, 2024. Estimating geographical retail markets from card spending data. In : *site de Economic Statistics Centre of Excellenc*e. [en ligne]. Novembre 2024. [Consulté le 29 janvier 2025]. Disponible à l'adresse : *https://www.escoe.ac.uk/publications/estimating-geographical-retail-markets-from-card-spending-data/*.

- INSEE, 2020. Éclairage - Disparités territoriales de consommation : que disent les données de transaction par carte bancaire ? In : *Notes et points de conjoncture de l'année 2020*. [en ligne]. 15 décembre 2020. [Consulté le 29 janvier 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4488582?sommaire=4473296&q=cartes+bancaires*.

- LECLAIR, Marie, 2019. Utiliser les données de caisses pour le calcul de l'indice des prix à la consommation. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. Insee. N° N3 , pp. 61-75. [Consulté le 29 janvier 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254225?sommaire=4254170*.

- OBSERVATOIRE DE LA SÉCURITÉ DES MOYENS DE PAIEMENT, 2024. Rapport annuel 2023. [en ligne]. Septembre 2024. In : *site de la Banque de France*. [Consulté le 29 janvier 2025]. Disponible à l'adresse : *https://www.banque-france.fr/system/files/2024-09/OSMP-2023.pdf*.

PRÉSENTATION DU NUMÉRO N13

Avec ce numéro N13, le Courrier des statistiques s'ouvre au-delà du système statistique public. Le premier article présente le pôle science des données de l'inspection générale des finances (IGF), qui intervient dans l'évaluation des politiques publiques. Son rôle est illustré par un exemple sur l'assurabilité des collectivités territoriales. Le papier suivant est consacré au pôle *data* de l'inspection générale des affaires sociales (Igas) : il réalise des analyses sur mesure dans les domaines du travail, de la santé et des solidarités, en mobilisant des données d'origines variées, des systèmes de gestion locaux au *web scraping*.

Le voyage se poursuit au cœur du système statistique public, avec la présentation du code officiel géographique (COG). Comme ses équivalents étrangers, il répertorie les territoires, des communes jusqu'aux pays, et leur attribue un code unique. Il sert pour le recensement de la population et alimente de très nombreuses bases administratives.

Enfin, un dossier présente les explorations menées par l'Insee de données détenues par les opérateurs privés. Le premier article dresse un panorama des différentes sources de données d'opérateurs privés déjà utilisées et les perspectives pour l'avenir, au regard des évolutions de la réglementation européenne. Le deuxième papier analyse le potentiel des données de téléphonie mobile pour l'étude des déplacements de population et les mécanismes de ségrégation spatiale. Enfin, le dernier papier expose les travaux menés à partir des données de transactions par carte bancaire CB et met en avant les usages possibles pour l'analyse conjoncturelle et l'étude des territoires.

LE TRI FACILE

ISSN 2107-0903
ISBN 978-2-11-162480-1

9 782111 624801

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE