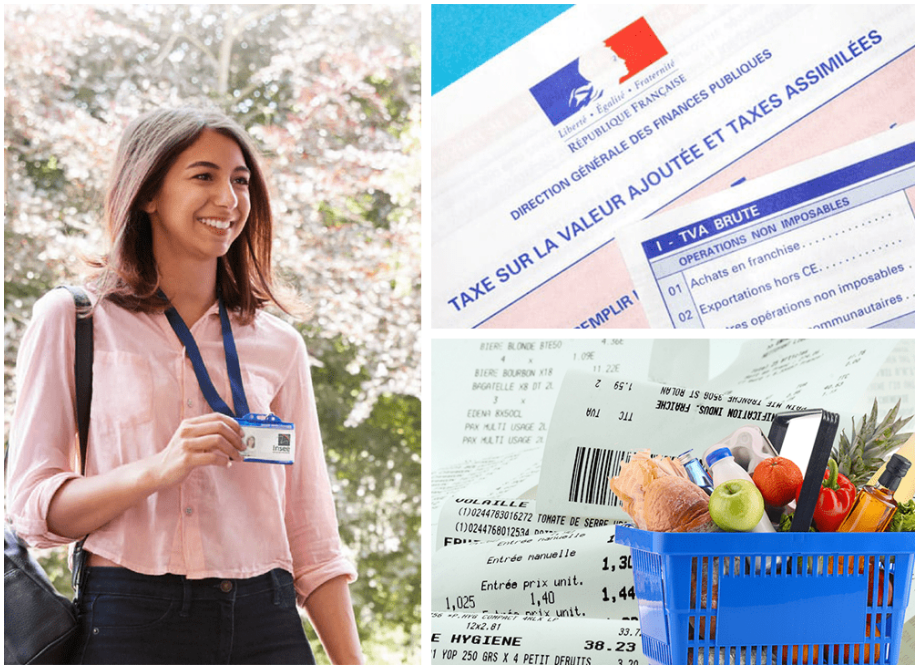
Quels types de sources l'Insee utilise-t-il pour construire ses statistiques ?

Publié le 15 mai 2023 sur le blog de l'Insee
Temps de lecture : 10 minutes

Françoise Dupont, Insee.



Pour pouvoir produire une statistique de qualité, encore faut-il savoir où chercher les informations ! Pour cela, les statisticiens de l'Insee utilisent une matière première riche et protéiforme, les données que les statisticiens vont puiser dans trois types de sources : la collecte directe, le réemploi de données d'origine administrative ou privée.

Les données susceptibles d'être utilisées par les statisticiens présentent des avantages et des inconvénients, car aucune d'entre elles n'est parfaitement ciblée ou exhaustive. Pour passer d'une donnée individuelle brute à un indicateur agrégé, il y a beaucoup de travail à fournir. C'est le cœur du métier des experts en statistiques de l'Insee. Or, comme le rôle de l'Insee consiste à fournir une réponse (sous la forme de statistiques) à une question précise, ce travail commence par l'identification de la ou des sources permettant de construire cette statistique avec justesse et précision. Parfois, il faut combiner des données de différentes origines pour produire des statistiques pertinentes.

Pour choisir les données qu'ils vont mobiliser, les statisticiens arbitrent selon la qualité des sources disponibles, mais aussi selon des considérations de budget, de temps nécessaire, de contrainte de délais, de type d'information souhaitée, de niveau de détail souhaité pour les statistiques à publier (Angel, 2023). Par exemple, la réalisation des enquêtes par sondage présente un coût, à la fois pour l'Insee qui doit mobiliser ses agents, mais aussi pour les ménages ou les entreprises enquêtés qui doivent dégager du temps pour répondre aux questions. L'Insee privilégie donc le recours aux données administratives lorsque c'est adapté et réserve les enquêtes pour recueillir des informations qui ne seront jamais disponibles dans des sources administratives (comme des ressentis sur le confort du

logement, des projets de consommation, des motivations de choix de transports pour des déplacements, etc.).

À quoi correspondent les trois types de sources disponibles ? Comment et quand l'Insee les mobilise-t-il ? Quels sont leurs avantages respectifs ? On fait le point.

La collecte directe : les enquêtes par échantillonnage

L'Insee réalise des enquêtes en interrogeant un échantillon représentatif de personnes ou d'entreprises sélectionnées au hasard (Ardilly et *alii*, 2022). Ces enquêtes sont réalisées via trois modes d'interrogation :

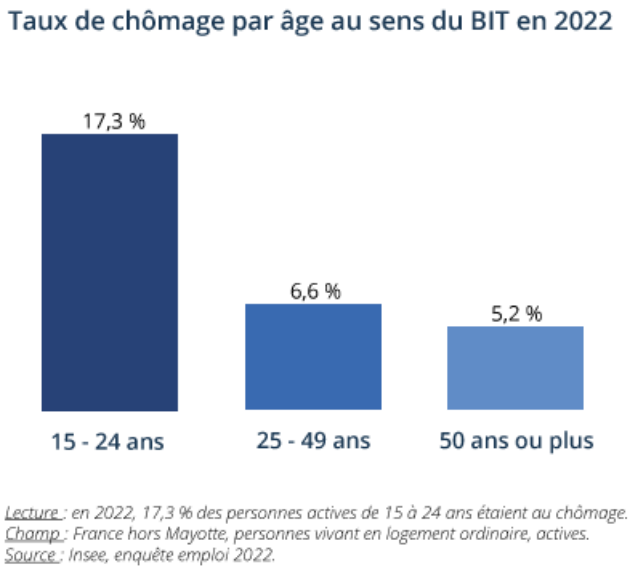
- la réponse à un questionnaire sur internet (ou imprimé) rempli directement par les personnes ou les entreprises ;
- la réponse à des questions posées par téléphone par un enquêteur ;
- la réponse à un questionnaire réalisé lors d'une visite au domicile (ou sur le lieu de l'entreprise) auprès des personnes (ou entreprises) enquêtées par des enquêteurs de l'Insee.

Parmi les statistiques obtenues par voie d'enquête, on retrouve le nombre d'habitants en France, les statistiques sur les conditions de vie, les chiffres du chômage ou encore certains chiffres sur le marché du travail (voir exemple ci-dessous), les prévisions d'activité dans les mois à venir des entreprises, les évolutions récentes et la structure de l'activité des entreprises.

Un exemple : l'enquête Emploi

L'enquête Emploi figure parmi les enquêtes les plus connues de l'Insee. Elle permet notamment de calculer le taux de chômage selon la définition du Bureau International du Travail (BIT), pour laquelle on a besoin de savoir si la personne est disponible pour occuper un emploi et si elle fait des démarches pour trouver un emploi (informations indisponibles dans des données administratives – *figure 1*). Menée sur un gros échantillon (environ 90 000 personnes de plus de 15 ans interrogées chaque trimestre), cette enquête permet à l'Insee de suivre chaque trimestre différents indicateurs du marché du travail en France. Il n'est pas nécessaire d'interroger chaque logement : un échantillon tiré au hasard suffit. Pour disposer de résultats représentatifs, il est important que le maximum de personnes sélectionnées réponde à l'enquête. Pour la première visite les personnes sont contactées par un enquêteur ou une enquêtrice de l'Insee, tenus au secret professionnel, en vue d'une interrogation en face-à-face. Les interrogations suivantes se font par téléphone avec une personne de l'Insee ou par Internet.

Figure 1 – Taux de chômage par âge au sens du BIT en 2022



Lecture : en 2022, 17,3 % des personnes actives de 15 à 24 ans étaient au chômage.
Champ : France hors Mayotte, personnes vivant en logement ordinaire, actives.
Source : Insee, enquête emploi 2022.

Les sources administratives

Ce sont les données déjà collectées auprès des personnes ou des entreprises via les différents services de l'administration (ou une personne morale de droit public, ou une personne morale de droit privé gérant un service public) pour répondre à des objectifs opérationnels.

En pratique, ces données proviennent d'administrations comme :

- La Direction générale des finances publiques pour des données de déclarations de revenus, de collecte d'impôts, ou de déclarations de TVA ;
- La Caisse nationale d'allocations familiales (CNAF) et la Mutualité sociale agricole (MSA) pour des données de versement des prestations familiales ou sociales ;
- Mais aussi… de la Caisse nationale d'assurance vieillesse (CNAV), l'Union de recouvrement des cotisations de sécurité sociale et d'allocations familiales (URSSAF)-caisse nationale, le Groupement d'intérêt public Modernisation des Déclarations Sociales ;
- etc.

Parmi ces données, conformément aux principes de nécessité et de minimisation, seules celles qui sont utiles à l'élaboration de statistiques peuvent être transmises à l'Insee. Elles sont alors communiquées par des moyens sécurisés et couvertes par le secret statistique.

Par comparaison avec les enquêtes par sondage, les sources administratives présentent quatre avantages. Elles permettent :

1. Un **recueil plus précis** pour les sujets complexes ou qui font appel à la mémoire. Les données proviennent de déclarations officielles et obligatoires, ce qui réduit donc les risques d'erreurs (durée de travail rémunérée, revenus…) ;

2. Un **allégement de charge** pour les personnes ou les entreprises enquêtées ;
3. Des **données plus fines** en particulier au niveau local grâce à l'exhaustivité du recueil, ce que ne permettent pas les enquêtes, dont la taille de l'échantillon n'autorise au mieux qu'une fourniture de statistiques au niveau régional ;
4. Un coût de collecte plus faible pour l'Insee que les enquêtes.

A contrario, elles présentent des inconvénients :

1. On ne dispose pas toujours du **concept** que l'on souhaite mesurer. Par exemple, le « chômage au sens du BIT » ne peut pas être reconstitué à partir des seules données administratives comme celles de Pôle emploi, dans la mesure où celles-ci ne renseignent pas sur la disponibilité pour travailler ni sur l'effectivité de la recherche d'emploi ;
2. La source peut être affectée par une **modification législative** (par exemple la suppression de la taxe d'habitation pour les résidences principales), ce qui génère des coûts d'adaptation et peut poser des difficultés de comparaison dans le temps ;
3. Elle ne couvre pas toujours l'ensemble des personnes que l'on veut étudier et il faut parfois la compléter par une autre source (par exemple la MSA en complément de la CNAF pour les prestations familiales et sociales) ;
4. Elle peut être affectée par des modifications dans les consignes d'enregistrement des actes à l'origine des données (par exemple : demandeurs d'emploi, dépôt de plaintes, dépôt de permis de construire, diagnostic de performance énergétique…).

Par ailleurs, l'utilisation d'une source administrative nécessite un investissement de départ. Il faut établir un lien de partenariat avec les fournisseurs pour mettre en place la livraison et le conventionnement et pour s'approprier les forces et faiblesses de la source. Il faut ensuite mettre au point les traitements.

Le fondement juridique de l'utilisation des données administratives à des fins de statistique
En France, un dispositif juridique encadre rigoureusement la mobilisation et la protection des données utilisées à des fins de statistique par l'Insee. Le recours aux sources administratives est possible grâce à l'introduction en 1986 de l'article 7 bis de la loi du 7 juin 1951 modifiée sur l'obligation, la coordination et le secret en matière de statistiques. Depuis 2004, la transmission de données administratives est obligatoire lorsqu'elle émane d'une demande du ministre chargé de l'économie établie après avis du Conseil national de l'information statistique (Cnis). Enfin, ces données sont placées sous la protection du Comité du secret statistique grâce à l'article 7 ter de la loi du 7 juin 1951.

Parmi les statistiques employant des données d'origine administrative, on retrouve les statistiques issues des bulletins d'état civil (naissances, décès), des statistiques concernant les résultats des entreprises (données fiscales, chiffres d'affaires via les déclarations de TVA), des données sur les revenus ou sur l'emploi salarié (cf. ci-dessous).

En pratique : les données sur l'emploi salarié et les salaires

La DSN (pour Déclaration Sociale Nominative) est une formalité déclarative que doivent accomplir les entreprises employant des salariés, pour les administrations sociales et fiscales. L'objectif de cette déclaration est double : satisfaire le plus grand nombre possible d'obligations déclaratives existant dans le champ social, attribuer des droits aux salariés. Elle est par ailleurs mobilisée à des fins statistiques. Ainsi, pour les salariés, l'Insee dispose de la nature des emplois (CDD, CDI, intérim, etc.), leur quotité de travail en cas de temps partiel, leur catégorie socio-professionnelle, leur lieu de travail, leurs autres emplois en cas de multi-activité ou encore leurs rémunérations *(figure 2)*. L'institut dispose également

d'informations concernant l'ancienneté du contrat ou des événements intervenant sur le contrat (début, rupture, interruption…). Ces données permettent donc de produire des statistiques sur les effectifs salariés, sur les salaires, sur les volumes de travail rémunéré, par catégories fines de secteur d'activité, de catégories professionnelles, de sexe, d'âge, … , et à un niveau géographique fin.

Figure 2 – Le contrat de travail, une notion centrale de la Déclaration Sociale Nominative

Les sources privées : les données des entreprises

L'apparition de données massives ("big data") issues pour l'essentiel d'entreprises privées ou publiques a entraîné des réflexions stratégiques et techniques de la part de l'Insee et des différents instituts statistiques nationaux et internationaux. Il peut s'agir des données de téléphonie mobile, des consommations relevées par les compteurs d'électricité ou de gaz, des données des transactions par carte bleue, ou encore des données des tickets de caisse des grandes enseignes de magasins.

L'Insee a mené des recherches pour évaluer le potentiel de certaines sources ainsi qu'une concertation sous l'égide du Cnis, avec les principaux détenteurs de sources. L'objectif ? Utiliser le potentiel des sources privées comme une opportunité permettant d'analyser certaines évolutions de façon très réactive et rapide qu'il n'est pas possible d'analyser avec des sources administratives ou des enquêtes.

À ce jour, une seule statistique produite régulièrement par l'Insee repose en partie sur des données massives d'origine privée. Il s'agit de l'indice des prix à la consommation. A titre exceptionnel, d'autres sources ont été mobilisées pour produire des statistiques pendant la crise sanitaire du Covid.

Le fondement juridique de l'utilisation des données privées à des fins de statistique
Depuis 2016, l'article 3 bis de la loi du 7 juin 1951 permet à la statistique publique d'avoir accès à des données présentes dans des bases de données détenues par des personnes morales de droit privé (entreprises privées, sociétés civiles, groupements d'intérêt économique, associations) pour des usages statistiques exclusivement dans le cadre d'enquêtes, et ce sous des conditions très strictes. À savoir :
– La décision doit être précédée d'une concertation avec les personnes de droit privé sollicitées et d'une étude de faisabilité et d'opportunité rendue publique ;
– Les données transmises par ces personnes ne peuvent faire l'objet d'aucune communication de la part du service dépositaire ;
– Les conditions dans lesquelles sont réalisées ces enquêtes sont fixées par voie réglementaire (faisabilité, opportunité, modalités de collecte, enregistrement, conservation et destruction).

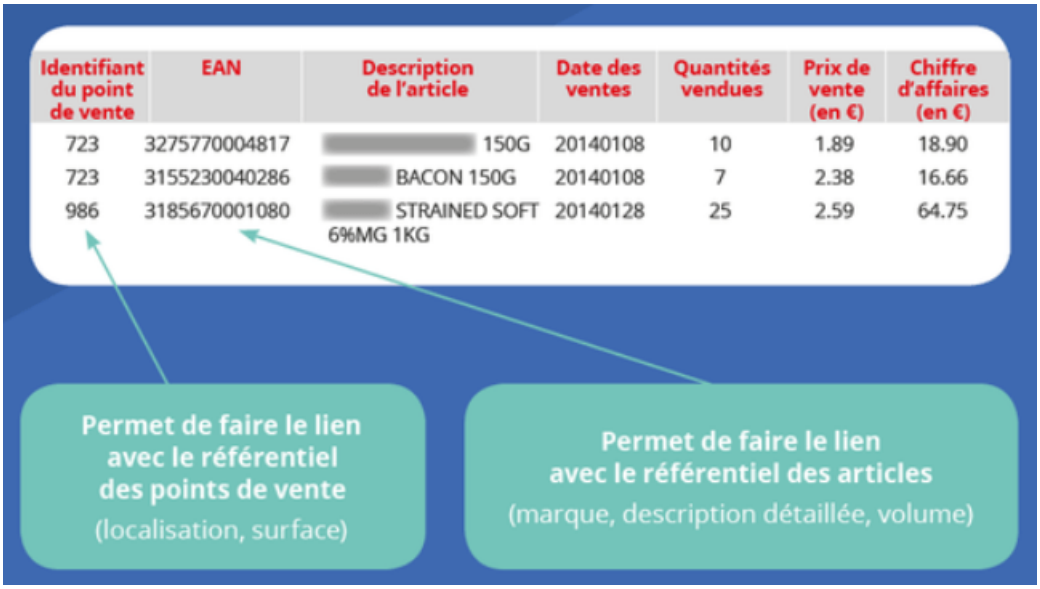
En pratique : l'indice des prix à la consommation

L'indice des prix à la consommation (IPC) constitue un exemple éclairant d'utilisation de cette méthode. Cet indice, qui permet de mesurer l'évolution générale des prix des biens et services que les Français consomment, se base sur l'observation du prix d'un panier moyen relevé à partir de trois sources :

- **Les enquêtes**. Des agents de l'Insee collectent directement et sur le terrain 150 000 prix de produits consommés par les ménages français sur l'ensemble du territoire et dans tous les types d'enseignes du commerce de proximité à la grande distribution ;
- **Les relevés de prix dématérialisés** issus de relevés sur internet ou de bases de données d'autres organismes (aviation civile pour les prix du transport aérien, caisse nationale de l'assurance maladie pour les services de santé…) ;
- **Les données de caisse**. Depuis 2020, les données des tickets de caisse, recueillies par les points de vente de la grande distribution lorsque le consommateur passe à la caisse du magasin, sont utilisées pour le calcul de l'indice des prix pour un ensemble de produits du quotidien : alimentaire industriel, entretien, hygiène, beauté *(figure 3)*.

En mobilisant ces différentes données, l'Insee peut ainsi bâtir un indice des prix à la consommation fiable, permettant de mesurer l'évolution générale des prix et d'analyser finement l'inflation (quand les prix augmentent) ou la déflation (quand les prix baissent).

Figure 3 – Un échantillon des données de caisse



Du recueil des données à la statistique, un important travail

Une fois collectée, cette matière première, quelle que soit son origine, doit être « raffinée » avant d'être agrégée pour aboutir à une statistique de qualité.

Il faut ainsi en contrôler la cohérence, identifier les anomalies et les corriger, compléter les réponses manquantes, corriger les éventuels défauts de représentativité, construire les concepts adaptés (l'emploi au sens du Bureau International du Travail, le niveau de vie…), modifier les données pour les mettre sous une forme adaptée pour l'analyse, codifier des libellés en clair dans une nomenclature (de professions et catégories socioprofessionnelles, de diplômes…), passer de données recueillies sur un échantillon à des données portant sur toute la population observée…

Ces traitements sont indispensables pour garantir la qualité des données produites ; ils sont documentés, ce qui permet aux utilisateurs d'en connaître les forces et faiblesses, et d'en déduire les conditions d'utilisation.

L'Insee utilise donc différents types de sources pour construire ses statistiques, en les combinant au mieux en fonction de leurs avantages et inconvénients. Ces pratiques ne sont pas figées et pour répondre aux questions qui leur sont adressées, les statisticiens continuent sans cesse d'explorer de nouvelles sources, ou de nouvelles techniques de collecte.

Ces pratiques ne sont pas propres à l'Insee ; les services statistiques de ministères utilisent aussi des sources diverses, notamment des enquêtes, des sources administratives ou des données collectées auprès d'entreprises ou d'opérateurs publics ; la part de ces différentes sources varie d'un ministère à l'autre, en fonction des questions posées, ou des données disponibles, avec un enjeu partagé : assurer la qualité de l'information produite. C'est le cas par exemple des statistiques produites par le service statistique ministériel de la sécurité intérieure, qui reposent sur deux piliers qui se complètent et s'enrichissent : les données administratives (infractions enregistrées par les services de gendarmerie et de police…) et les résultats des enquêtes de « victimation » auprès de la population.

Pour en savoir plus :

Le cadre général

- Insee, 2022, « Pour comprendre les statistiques de l'Insee » (vidéo – 3:35 min)

Les sources

- Angel J-W., 2023, « Elle est pas fraîche ma statistique ? », *Blog de l'Insee*, avril
- Cnis, 2006, « Enquêtes statistiques et sources administratives : une complémentarité à exploiter », *Chroniques du Conseil National de l'Information Statistique* n°5 (pdf – 4 Mo), décembre

Les enquêtes

- Ardilly P., Castell L. et Sillard P., 2022, « Il y a sondage et sondage… », *Blog de l'Insee*, juillet

Les données d'origine privées

- Leclair M., 2019, « Utiliser les données de caisses pour le calcul de l'indice des prix à la consommation », *Courrier des statistiques* n°3, décembre
- Ourliac B., 2020, « Google en sait-il plus que l'Insee sur les Français ? », *Blog de l'Insee*, décembre

Le cas particulier de la crise sanitaire : un terrain d'innovation et de progrès pour de nouvelles données et de nouvelles méthodes

- Tavernier J-L., 2020, « La statistique à l'épreuve de la crise sanitaire », *Blog de l'Insee*, mai
- Bureau B., Duquerroy A., Giorgi J., Lé M., Scott S. et Vinas F., 2021, « Crise sanitaire : une approche complémentaire sur l'activité des entreprises », *Blog de l'Insee*, avril
- Coudin É., Galiana L., Poulhes M., Semecurbe F., et Suarez Castillo M., 2020, « Mesure de population présente en temps de confinement et statistiques expérimentales », *Blog de l'Insee*, avril
- Houriez G., 2020, « Santé, éducation, services administratifs : la difficile mesure des activités non marchandes en temps de crise sanitaire », *Blog de l'Insee*, novembre
- Pouget J., 2021, « L'arithmétique de la crise : bien comprendre les chiffres de croissance en temps de Covid-19 », *Blog de l'Insee*, octobre
- Pouget J., 2020, « Nouvelles données pour suivre la conjoncture économique pendant la crise sanitaire : quelles avancées ? quelles suites ? », *Blog de l'Insee*, juillet