« Elle est pas fraîche ma statistique ? »

Où le statisticien opère le délicat arbitrage délai-qualité… et coût

Publié le 21 avril 2023 sur le blog de l'Insee
Temps de lecture : 12 minutes

Jean-William Angel, Insee.

©John Tenniel

Les statisticiens ne prennent pas leur temps. Ils ont démontré pendant la crise sanitaire qu'ils savaient réagir et produire vite : en exploitant des données massives et à haute fréquence, en lançant en quelques semaines des enquêtes inédites, en adaptant leurs méthodes. En temps normal aussi, ils s'appliquent à réduire leurs délais de production : 15 jours ont été gagnés sur le calcul de l'évolution trimestrielle du produit intérieur brut (PIB) et du taux de chômage depuis 2016 ; 5 sur la production industrielle depuis 2022 ; un indice des prix provisoire est publié avant la fin du mois. Toutefois, le temps gagné se paie parfois : par des données moins précises ou moins détaillées.
Les statisticiens ne prennent pas leur temps : ils sont soumis à celui des informations, qu'ils recueillent dans les fichiers administratifs ou dans les enquêtes qu'ils mènent. Les premières sont souvent exhaustives mais ne sont pas rapidement disponibles et nécessitent des travaux de retraitement et d'appariement complexes. Les enquêtes sont strictement encadrées et élaborées. Elles ont aussi un coût, plusieurs millions d'euros pour les plus grandes, et font peser une charge sur les répondants que la statistique publique s'efforce de minimiser.
Les statisticiens ne prennent pas leur temps, ils arbitrent. Entre fraîcheur des données, précision des résultats et coût pour la nation.

« On est en 2023 et l'Insee fournit seulement les statistiques détaillées du recensement de 2019 ! » « La dernière enquête sur les sans-domicile date d'il y a plus de dix ans ! » « Il faudra attendre la fin 2023 pour avoir une vision complète du système productif en 2021 : qu'est-ce que vous voulez faire avec ça ? »
De telles phrases fleurissent de temps en temps sur les réseaux sociaux ou dans les médias.

À l'heure des données massives (big data) et des réponses instantanées des moteurs de recherche à toutes nos requêtes, ces remarques peuvent paraître de bon sens. Il suffirait de se baisser pour récolter des statistiques toutes fraîches. Les statisticiens mettraient-ils alors de la mauvaise volonté à utiliser ces données, préférant conduire des enquêtes à leur façon ? Seraient-ils passés à côté des derniers développements technologiques ? En un mot, leur logiciel est-il adapté aux réalités d'aujourd'hui ?

Toutes les critiques sur « le temps long de la statistique » ne sont pas infondées. Néanmoins, la plupart relèvent d'une méconnaissance de la pratique du statisticien. Pour mieux les apprécier, il faut ouvrir le capot pour comprendre d'où vient que ce temps est parfois (relativement) long et pourquoi il est utile de le prendre. Mais commençons par retracer les efforts que la statistique publique, au premier rang de laquelle l'Insee, a su déployer lors de la crise de la Covid pour être aussi réactive que possible.

Action, réaction, adaptation

Le premier confinement a débuté le 17 mars 2020. Le 26, l'Insee produisait une première estimation de la chute de l'activité et de celle de la consommation. Puis, deux fois par mois, l'institut allait publier un état simplifié de la situation économique.
Cette réactivité, l'institut la doit à sa capacité d'adaptation. Pour assurer la continuité de leurs missions, les statisticiens ont modifié certaines enquêtes auprès des ménages, dont la collecte est passée du face-à-face aux entretiens téléphoniques ; ils ont utilisé de nouvelles sources de données disponibles instantanément – telles que les données des téléphones mobiles ou de transactions par carte bancaire ; et ils ont mobilisé de nouvelles méthodes – principalement le *nowcasting,* littéralement prévision du présent.
Dans cette situation inédite et complexe, la statistique publique a donc été capable de compresser le temps nécessaire à « mesurer pour comprendre ». Comprendre, en temps réel, quelle était la situation économique française, où était localisée la population sur l'ensemble du territoire national, et comment évoluait la mortalité.

Pour suivre au plus près l'activité économique, il fallait recourir à de nouvelles sources. Celles-ci sont dites à haute fréquence car elles sont disponibles au jour le jour, voire plus fréquemment. En la matière, les conjoncturistes font feu de tout bois. Ils se penchent sur les statistiques de requêtes dans les moteurs de recherche, les données de transport ferroviaire et de trafic routier, ou encore celles de consommation d'électricité. Ils testent également d'autres pistes, comme les indicateurs de pollution ou encore le vocabulaire utilisé dans la presse. Avec ce dernier, l'Insee construit un « indice de sentiment médiatique », qui aura permis de conforter la chute de l'activité. Mais l'Insee privilégie aussi les statistiques issues des transactions par cartes bancaires, obtenues grâce à un accord inédit avec le Groupement des Cartes bancaires CB.
Enfin, l'Insee met les bouchées doubles pour évaluer l'activité des entreprises selon les secteurs. D'une part, en exploitant les toutes récentes déclarations sociales nominatives (DSN), que ce soit pour suivre les heures rémunérées, indicateur précieux de l'activité, ou pour mesurer l'emploi. D'autre part, en estimant le chiffre d'affaires perdu à la suite du choc sur l'activité, en comparant les déclarations de TVA à ce qu'aurait été la situation en l'absence de crise sanitaire [Bureau et *alii*, 2021]. Ces travaux, menés avec la Banque de France, sont complétés par une évaluation des chocs de trésorerie.

La répartition des Français sur le territoire national est une information capitale pour les services de santé, d'approvisionnement et de police. L'Insee la publie dès le 8 avril 2020. Là encore, ce travail est rendu possible grâce à une collaboration, hélas seulement ponctuelle, avec l'opérateur Orange.
Quant aux statistiques de décès quotidiens, elles paraissent chaque semaine dès le 27 mars avec une fraîcheur inédite : jusqu'à J-11 pour l'ensemble des communes et jusqu'à J-7 pour les communes qui transmettent les actes de décès sous forme dématérialisée. Elles sont de surcroît accompagnées de commentaires pour interpréter la comparaison avec les années antérieures.

Enquêtes inédites

Aussi précieuses soient ces nouvelles sources, elles ne remplacent pas la matière première du statisticien : les enquêtes. Nous verrons plus loin qu'il faut du temps pour les construire, les administrer et les exploiter. Toutefois, en pleine crise sanitaire, la statistique publique a su, en un temps record, adapter ses sources existantes, voire bâtir des interrogations indispensables pour les compléter. Qu'on en juge :
– adaptation d'une enquête pour mesurer le recours au télétravail et au chômage partiel par la Dares, le service statistique du ministère du Travail, avec l'appui de l'Insee. Les résultats sont publiés dans la foulée de l'enquête, le 17 avril 2020 ;
– élaboration et lancement en mai 2020 de la première enquête EpiCov, très grande enquête internet en partenariat avec l'Inserm et la Drees, service statistique du ministère de la Santé, pour prendre la mesure de la prévalence du virus et de ses symptômes ;
– questions spécifiques aux ménages dans l'enquête mensuelle sur leur confiance dans l'économie, afin d'apprécier les effets du confinement sur leur vie. La première exploitation sort en juin 2020 ;
– analyse textuelle des réponses des entreprises aux questions ouvertes pour prendre la mesure de l'inquiétude générale suscitée par l'épidémie (juillet 2020).

Tout récemment encore, fin 2022, l'Insee a adapté en un mois ses enquêtes de conjoncture auprès des entreprises pour éclairer l'impact de la crise énergétique. En parallèle, il a déployé une enquête spécifique auprès des fournisseurs d'électricité pour estimer l'évolution des tarifs qui seront facturés aux professionnels en 2023.

Gagner sur les délais des indicateurs conjoncturels sans sacrifier la qualité

Les avancées qui précèdent portent sur le suivi de la conjoncture et ses indicateurs. La crise sanitaire les a commandées : il était crucial de rendre compte aussi vite que possible de la situation économique et sociale. Mais, heureusement, les statisticiens n'attendent pas les crises pour livrer bataille contre le temps. C'est d'ailleurs parce qu'ils y travaillaient depuis plusieurs années, qu'ils ont pu mobiliser ces ressources lors de la crise sanitaire. Mieux explorer les possibilités des données à haute fréquence figurait dès 2016 parmi les recommandations du rapport Bean, consacré à l'évaluation des statistiques économiques officielles britanniques [Blanchet et Givord, 2017].

En temps normal aussi, diminuer les délais de parution des indicateurs conjoncturels est un impératif. Il faut à la fois satisfaire les utilisateurs et respecter des règlements européens, tout en veillant à préserver la qualité des chiffres.
Depuis janvier 2016, l'Insee a ainsi avancé de 15 jours la date de publication de sa première estimation de la croissance trimestrielle du produit intérieur brut (PIB) en volume. Celle-ci est désormais diffusée 30 jours après la fin du trimestre considéré.
Privilégier le délai d'information des utilisateurs conduit à restreindre le nombre de données sur lesquelles reposent les calculs, ou le temps consacré à leur vérification. *De facto*, l'estimation à 30 jours est potentiellement plus sujette à révision que celle à 45. Aussi, avant de décider la diffusion à la nouvelle échéance, les statisticiens ont-ils calculé l'écart sur la croissance entre les tests effectués à 30 jours et les données publiées à 45 jours. Le feu vert a été donné car cet écart s'est avéré très limité, de l'ordre de 0,07 point de PIB.
En parallèle, d'autres délais ont été réduits, comme celui de l'indice de la production industrielle (IPI), publié mensuellement. Cinq jours ont été gagnés depuis 2022, l'IPI étant désormais publié 35 jours après la fin du mois concerné. L'objectif à terme étant de publier à 30 jours.
L'année où le PIB était avancé de 15 jours, le taux de chômage l'était aussi, avec les autres indicateurs trimestriels sur le marché du travail que fournit l'enquête Emploi. Grâce à l'exploitation anticipée de la DSN, le délai de diffusion de l'estimation flash de l'emploi salarié est passé de T+45 jours avant crise à

T+35 jours environ aujourd'hui.
Du côté des prix à la consommation, un indice provisoire est aujourd'hui produit juste avant la fin du mois. Il est la plupart du temps confirmé par l'indice définitif, publié au milieu du mois suivant. Bien sûr, le degré de détail n'est pas le même : l'utilisateur doit accepter de payer ce prix pour un délai raccourci.

Terminons ce panorama des sources infra-annuelles avec les données de chiffre d'affaires, issues des déclarations de TVA que les entreprises remplissent chaque mois. Celles-ci sont exploitées de façon exhaustive et permettent de publier des indices au niveau le plus fin de la nomenclature 60 jours après la fin du mois. Nous avons vu plus haut que ces données de TVA avaient été précieuses en temps de crise sanitaire.

Le recueil des données dicte les délais pour certains indicateurs annuels

Les statistiques illustrant la situation socio-économique d'une année donnée, notamment la distribution des revenus, ne sont pas soumises aux mêmes impératifs et ne bénéficient pas des mêmes souplesses que les indicateurs conjoncturels. Les données administratives sur lesquelles elles s'appuient ne sont pas disponibles très rapidement, les questions de leurs enquêtes nécessitent que les entreprises ou les ménages prennent du temps pour y répondre. Néanmoins, les statisticiens travaillent aussi à réduire les délais dont ils ont la maîtrise. Le taux de pauvreté est un bon exemple de leurs efforts, tout autant que de leurs difficultés. Car ils se heurtent parfois à un temps incompressible.

Pour calculer le taux de pauvreté monétaire, il faut connaître la « distribution des niveaux de vie », et pour cela l'ensemble des revenus de chaque personne, ou d'un échantillon représentatif de l'ensemble de la population. De cette distribution, on déduira la proportion de gens dont le niveau de vie les place sous le seuil de pauvreté, soit 60 % du niveau de vie médian de la population.
L'ensemble des revenus couvre les prestations et minima sociaux, les revenus d'activité (salaires, revenus des indépendants), de remplacement (retraite, chômage, etc.) et du patrimoine (foncier et financier). Pour connaître les niveaux de vie, il faudra ensuite retirer les impôts directs.
Presque deux ans sont nécessaires pour produire les données au niveau de robustesse et de détail requis : les trois quarts environ sont dévolus au recueil des données des administrations fiscales et des caisses de sécurité sociale, et à leur transmission à l'Insee ; un quart est utilisé par les statisticiens pour leurs travaux. Il faut en effet attendre la fin de l'année n+1 pour obtenir la totalité des déclarations fiscales des ménages et donc les fichiers fiscaux de l'année n, le temps de traiter les dossiers les plus complexes (changements de foyers fiscaux, déclarations modificatives…). Quant aux derniers fichiers sociaux, ils sont transmis au printemps n+2. S'ensuit un lourd travail d'appariement avec l'enquête Emploi et de complétion des données manquantes. Les résultats de ce dispositif, appelé ERFS pour enquête sur les revenus fiscaux et sociaux, sont publiés à l'automne n+2.

Afin de livrer ces chiffres plus rapidement, l'Insee a décidé il y a quelques années d'estimer des indicateurs avancés du taux de pauvreté et des inégalités de niveaux de vie, une dizaine de mois après la fin de l'année. Pour y arriver, les statisticiens « prévoient le présent » en utilisant la microsimulation et s'appuient sur les informations disponibles quelques mois après la fin de l'année observée.
Ils gagnent ce faisant une année sur les calculs issus d'ERFS. Et, comme à chaque fois, le raccourcissement des délais s'accompagne d'une moindre précision ou d'informations moins détaillées. Ainsi, cette simulation porte sur l'évolution du taux de pauvreté et des indicateurs d'inégalités, mais ne peut étudier en détail l'évolution de tous les types de revenus, le long de l'ensemble de l'échelle des niveaux de vie.

La lourdeur, qui va de pair avec la fiabilité de la mécanique ERFS, tient à ce qu'elle combine données administratives (fiscales et sociales) et enquête statistique (Emploi). Outre qu'elles sont parfois longues

à obtenir, les premières n'ont, par définition, pas été conçues pour répondre à des besoins statistiques. Faudrait-il alors ne passer que par les enquêtes statistiques ?

Des enquêtes pour mesurer des réalités complexes, dans un cadre strict

En l'espèce, le recours aux données administratives fiscales garantit une information de bien meilleure qualité que les déclarations de revenus faites par les personnes interrogées par enquête. Mais c'est surtout le coût financier, pour la nation, et la charge pour les répondants, ménages ou entreprises, qui conduisent depuis longtemps l'Insee et la statistique publique à concentrer le recours aux enquêtes sur ce qui ne se trouve pas dans les données administratives. À titre d'exemple, l'enquête Emploi, la plus grande après l'enquête annuelle de recensement, coûte 15 millions d'euros par an, emploie quelque 130 enquêteurs en équivalent temps plein et interroge 90 000 personnes chaque trimestre.
L'enquête Emploi relève d'un règlement européen et les indicateurs qu'elle produit sont cruciaux pour les décideurs économiques et politiques, ainsi que pour l'information du citoyen. Le suivi trimestriel est justifié par le caractère conjoncturel des fluctuations du marché du travail. Son questionnement permet en particulier de recueillir des informations essentielles pour la mesure du chômage, qu'on ne trouve pas ailleurs : avoir effectué des démarches actives pour trouver un emploi ou être disponible pour travailler. C'est là tout l'intérêt des enquêtes de la statistique publique, tant auprès des ménages que des entreprises : de pouvoir mesurer des comportements et des réalités complexes que ne permettent pas d'appréhender les sources administratives.
Ces enquêtes sont établies selon des normes rigoureuses, nécessitent des tests approfondis, demandent une concertation élargie et sont validées par un label.

Il en va ainsi par exemple de la prochaine enquête sur les personnes sans-domicile. Cette enquête est dite structurelle en ce sens que l'image de la réalité sociale ou économique qu'elle délivre ne varie pas tant qu'il faille la répéter chaque année pour porter un diagnostic fiable. Il s'agit de recenser le nombre de personnes sans-domicile, mais aussi de connaître leur situation et leur trajectoire. Comme toutes les enquêtes de la statistique publique, elle fait l'objet d'un avis du Conseil national de l'information statistique, le Cnis. Chargé de la concertation entre producteurs et utilisateurs de la statistique publique, il délivre l'avis d'opportunité au terme d'un examen attentif. Cet avis atteste que l'enquête est bien une enquête statistique, qu'elle correspond à un besoin d'intérêt public et ne fait pas double emploi avec d'autres sources déjà disponibles (enquête statistique ou administrative, fichier de gestion, etc.). Puis, le Comité du label de la statistique publique s'assure, notamment, que l'enquête répond aux critères de qualité statistique, de pertinence du questionnement et qu'elle n'entraîne pas de charge excessive sur les enquêtés. Il attribue alors le cas échéant un label d'intérêt général et de qualité statistique, assorti éventuellement d'une proposition d'octroi de l'obligation de réponse.

Le processus de validation de l'enquête et de son protocole de collecte peut certes sembler lourd et long à mettre en œuvre. Mais c'est un gage de qualité d'autant plus nécessaire que ces enquêtes fournissent des éléments de cadrage qui permettent de corriger les biais de sources alternatives, disponibles rapidement mais incomplètes.

Le prix de la fraîcheur

Ainsi, sauf en cas de crise, la statistique publique ne décide pas du jour au lendemain de lancer une enquête comme on lance un sondage d'opinion [Ardilly et alii, 2022]. En particulier, parce qu'il s'agit de couvrir toute la population et pas seulement la plus facile à contacter, ce qui nécessite de déployer beaucoup d'efforts. Ici aussi les budgets sont donc élevés : 8,5 millions d'euros pour l'enquête sans-domicile de 2025. Celle-ci mobilisera plusieurs centaines d'enquêteurs pendant plusieurs mois et sera précédée en 2024 d'une enquête préparatoire auprès des structures d'aide et d'hébergement. L'Insee prévoit d'interroger environ 15 000 personnes.

À quelle fréquence la nation est-elle prête à payer un tel prix ? Comment arbitrer entre des clichés rapprochés, mais coûteux, et des photos plus espacées, quitte à ne pas voir évoluer assez finement le visage de la société ?
Cette question a été tranchée de façon originale pour le recensement de la population. Exhaustif jusqu'en 1999, cette opération colossale s'espaçait de plus en plus, aussi pour des raisons budgétaires. Neuf années s'étaient écoulées depuis le précédent : les Contrats de Plan État Région du début des années 2000 avaient été définis sur la base de données datant de 1990. L'Insee a donc choisi de changer ses méthodes et procède depuis 2004 à des photos annuelles sur un « échantillon ». Le terme ne doit pas tromper : cet échantillon couvre chaque année plus de 9 millions d'habitants.
Résultat : pour un coût global équivalent sur sept ans à celui d'un recensement exhaustif, le pays dispose chaque année d'une photo, « médiane » des années récentes. En décembre 2022 ont été diffusés les chiffres de la population 2020 de chaque commune ; en juin 2023, seront diffusés les résultats statistiques détaillés de 2020, qui prennent en compte les recensements de 2017 à 20…22 ! Une réflexion est en cours pour raccourcir ces délais, encore une fois en s'appuyant plus largement sur les données administratives. Sans compter que le recensement permet d'estimer dès la mi-janvier la population globale de notre pays au 1er janvier, publiée dans le bilan démographique.

Alors, elle est toujours pas fraîche ma statistique ?

Pour en savoir plus

- Tavernier, J.-L., 2020, « La statistique publique à l'épreuve de la crise sanitaire », *Blog de l'Insee*, mai
- Albouy V. et Legleye S., 2020, « Conditions de vie pendant le confinement : des écarts selon le niveau de vie et la catégorie socioprofessionnelle », *Insee Focus* n°197, juin
- Ardilly P., Castell L. et Sillard P., 2022, « Il y a sondage et sondage… », *Blog de l'Insee*, juillet
- Blanchet D. et Givord P., 2017, « Données massives, statistique publique et mesure de l'économie », in : *L'Économie française, édition 2017,* [en ligne]. *Insee Références*, pp. 59-77
- Blasco J., 2019, « Pourquoi les études de l'Insee sur les revenus ont deux ans « de retard » ? », *Blog d'Alternatives économiques*, décembre
- Bureau B., Duquerroy A., Lé M. et Vinas F. (Banque de France), Giorgi J. et Scott S. (Insee), 2021, « Crise sanitaire : des chocs de trésorerie (très) hétérogènes », *Blog de l'Insee*, juillet
- Bureau B., Duquerroy A., Lé M. et Vinas F. (Banque de France), Giorgi J. et Scott S. (Insee), 2021, « Crise sanitaire : une approche complémentaire sur l'activité des entreprises », *Blog de l'Insee*, avril
- Papon S., 2023, « Bilan démographique 2022 – L'espérance de vie stagne en 2022 et reste inférieure à celle de 2019 », *Insee première* n°1935, janvier
- Insee, 2023, « Nombre de décès quotidiens France, régions et départements », *Chiffres détaillés*, avril
- Insee, 2020, « Fonctionnement de l'Insee dans la période de confinement », *Courrier des statistiques* N5, décembre
- Insee, 2020, « Point de conjoncture du 26 mars 2020 »
- Insee, 2020, « Population présente sur le territoire avant et après le début du confinement – Premiers résultats », *Communiqué de presse*, avril
- Insee, 2016, « Indicateurs avancés, estimations précoces, nowcasting : où en est l'Insee ? », *Dossier de presse*