À propos du temps de production des statistiques publiques

Pierre Lamarche[*] *et Pascal Rivière*[**]

Si la production d'informations conjoncturelles s'opère dans des délais très brefs, nombre de statistiques publiques requièrent un temps significatif. Qu'elles proviennent d'enquêtes auprès de ménages ou d'entreprises, ou de sources administratives ou privées, trois grandes phases se dégagent. D'abord celle de la préparation : un temps de travail en amont, par exemple de concertation, qui n'entre pas dans le « délai de diffusion » des chiffres. Puis celle de la construction des données : via un processus de collecte pour les enquêtes, ou de transformation dans le cas de sources externes. Enfin, une dernière phase de traitement statistique et de diffusion, qui inclut en particulier des activités de validation.

Les phases de préparation et de traitement statistique sont d'autant plus chronophages que le nombre de variables est élevé. Le temps de la préparation dépend aussi fortement de la complexité et de la sensibilité du dispositif à établir. Celui de la construction croît généralement avec le nombre d'unités. Enfin, le temps passé sur la dernière phase dépend en particulier du niveau de détail auquel les statistiques sont diffusées.

De manière générale, la statistique publique produit certes des données, mais surtout du sens. Ainsi, même si une meilleure automatisation est envisageable, une large part du temps de travail réside dans l'analyse, l'interprétation, les vérifications, et toutes les boucles de rétroaction que cela induit.

Even though short-term business statistics can be produced very quickly, many official statistics require a significative amount of time. Be they originating in surveys towards households or businesses, or in administrative or private data, three main stages appear. First, preparation: a work period beforehand, dedicated to consultations for instance, that is not accounted for in the timeliness of statistics. Then, data construction: through a collecting process for surveys or a transformation one for external sources. Finally, data processing and distribution, which includes validation works in particular.

Preparation and data processing stages are more time-consuming the higher the amount of variables. Preparation time also depends greatly on the complexity and sensitivity of the process to devise. The construction time usually increases with the amount of units. Finally, time spent on the last stage depends specifically on the statistical granularity.

Generally speaking, statistical institutes produce insights as well as numbers, they also construct meaning and interpretation. Thus, even if a better automation may be worth considering, a large amount of the time spent on production resides in analysing, interpreting, verifying, and all the feedback loops that ensue.

* Chef de l'unité Innovation et stratégie du système d'information, DSI, Insee.
pierre.lamarche@insee.fr

** Chef de l'Inspection générale, Insee.
pascal.riviere@insee.fr

Dans un monde où les données sont abondantes et accessibles, dans un contexte d'essor du big data et de l'open data, mais aussi de possibilités d'appariement accrues (Benichou et al., 2023), un nouveau champ des possibles s'ouvre pour la statistique publique (Elbaum, 2018). Des plateformes innovantes de traitement des données rendent tout le processus de production statistique plus adaptable à de gros volumes et à des configurations hétérogènes (Comte et al., 2022). Elles le rendent aussi plus réplicable, avec la logique des *pipelines*[1] (Cotton et Haag, 2023). On observe également une prolifération d'outils facilitant le développement, entre autres via une utilisation croissante de l'intelligence artificielle.

On pourrait en déduire hâtivement que, quelle que soit l'origine des données, les statistiques publiques deviennent de plus en plus « faciles à produire » et que, profitant des nouvelles sources et de la *data science*, on devrait pouvoir désormais les réaliser beaucoup plus rapidement. Cette volonté de réduire les délais de production est d'ailleurs régulièrement affichée comme un objectif fort au niveau d'Eurostat[2].

> ***Nouvelles données ou pas, nouvelles techniques ou pas, l'élaboration de statistiques ayant une qualité suffisante pour servir de référence dans le débat public requiert toute une gamme d'opérations.***

Dans de nombreux cas, la rapidité de réalisation des statistiques est inhérente au processus, lorsque ce dernier a justement été conçu avec cette exigence principale (indice de production industrielle ou enquêtes de conjoncture, par exemple). Dans d'autres cas, des gains de temps ont effectivement été observés ces dernières années (statistiques sur l'emploi ou comptabilité nationale, par exemple). Pourtant, dans beaucoup de situations, il y a loin de la coupe aux lèvres. Nouvelles données ou pas, nouvelles techniques ou pas, l'élaboration de statistiques ayant une qualité suffisante pour servir de référence dans le débat public requiert toute une gamme d'opérations. Or, une grande partie d'entre elles ne se prête pas à une quelconque automatisation et leur accélération paraît de fait moins évidente.

L'objet du présent papier est de mieux comprendre tous les éléments qui expliquent le temps de production des statistiques publiques, et de déterminer ainsi les marges de manœuvre pour améliorer cet aspect sans dégrader les autres critères.

• Les marges de manœuvre peuvent différer selon les dispositifs mis en place. On distinguera ainsi **trois types de processus**[3] : les enquêtes, piliers historiques des statistiques officielles ; les processus fondés sur l'usage de données « administratives » (appartenant à des administrations, par exemple des organismes de protection sociale) ; enfin, ceux qui s'appuient sur des données « privées » (appartenant[4] à des entreprises privées, par exemple à des opérateurs de téléphonie mobile[5]). Il arrive, comme dans

1 Traitements de données standardisés.
2 Eurostat est l'institut statistique communautaire, direction générale de la Commission européenne.
3 Il existe aussi des situations hybrides, par exemple le cas des données relatives aux droits de mutation à titre gratuit (où il y a historiquement un échantillonnage, mais qu'on ne maîtrise pas).
4 Avec « appartenant », on veut simplement dire que ces données figurent dans des bases de données qui sont sous la responsabilité de ces entreprises (ou, respectivement, de ces administrations).
5 Lesur (2025) décrit les différentes sources privées qui ont été mobilisées par l'Insee dans le cadre de partenariats.

le cas des statistiques structurelles d'entreprises, que la production statistique mixe plusieurs types de processus.

• Le processus de production lui-même, de façon générale, peut être subdivisé en **trois grandes phases** : préparation, construction des données, traitement statistique et diffusion.

Pour chacune de ces phases, on analysera les déterminants du temps de réalisation des statistiques selon le type de dispositif. On verra également que la situation n'est pas la même selon que l'on diffuse des statistiques pérennes, régulières, ou au contraire ponctuelles, *one-shot*. On en tirera quelques leçons sur les principaux facteurs explicatifs du temps de réalisation, et sur les possibilités de gain de temps.

Mais avant cela, il nous faudra répondre à une autre question, a priori plus basique : au fait, de quel « temps » parle-t-on exactement ?

▶ Le temps nécessaire : plusieurs définitions, plusieurs moments-clés

Dans le processus complexe qui mène à la publication de statistiques sur un sujet, on peut distinguer quatre dates essentielles, qu'on notera de t1 à t4 (***figure 1***) :

• t1 : le démarrage de la réflexion – qui va se matérialiser, par exemple, par l'explicitation du besoin et le recensement des sources d'information disponibles ;

• t2 : la date de référence des statistiques – par exemple, on récupère des données fiscales portant sur le mois de décembre 2024, ou bien, on réalise une enquête auprès d'entreprises relativement à l'année 2024[6] ;

• t3 : la date de fin de récupération des données – c'est-à-dire la fin de collecte, pour une enquête, ou la date d'obtention de la dernière version du fichier administratif ;

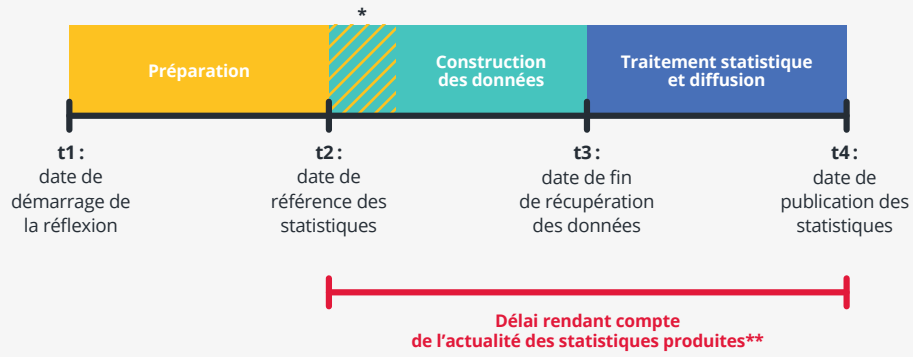
• t4 : la date de publication des statistiques.

Le critère privilégié pour apprécier l'actualité (ou *timeliness*) des statistiques produites, en référence au treizième principe du Code de bonnes pratiques de la statistique européenne[7], est le **délai entre la fin de la période de référence d'un phénomène statistique et la date de publication des résultats**. C'est donc t4 − t2.

Cette mesure, dont on comprend bien l'enjeu quant à la « fraîcheur » des données, ne rend pas compte de l'entièreté du temps de production statistique. En effet, elle fait l'impasse sur tout le travail effectué avant la période de référence, donc sur l'écart t2 − t1. Ainsi, pour l'indice des prix à la consommation, il a fallu une dizaine d'années pour intégrer des données de caisse[8] dans la production, entre la mise en place des conventions avec les enseignes de la grande distribution alimentaire et l'investissement

6 Lorsque la référence temporelle est en réalité une période, la date t2 que l'on considère est la fin de la période, donc le 31 décembre dans le cas d'une année, le dernier jour du mois dans le cas d'un mois. Pour la définition de la période de référence, voir le lien suivant : *https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Reference_period*.
7 Voir le lien : *https://www.insee.fr/fr/information/4140105*.
8 Voir le lien : *https://www.insee.fr/fr/metadonnees/definition/c2159*.

Figure 1 - Les trois phases du processus de production des statistiques publiques



* Pour certains processus, la fin de la préparation / le début de la collecte peut être postérieur à la date de référence des statistiques. C'est le cas par exemple pour les données fiscales, les déclarations de revenus fiscaux étant postérieures à la période de référence.
** L'actualité (*timeliness* en anglais) est un critère de qualité des statistiques publiques mis en avant dans le Code de bonnes pratiques de la statistique européenne.

informatique requis (Leclair, 2019). Bien sûr, ce temps d'investissement n'est plus nécessaire pour les périodes suivantes (ici, les mois suivants), mais on ne peut passer sous silence cette longue préparation.

Par ailleurs, lorsqu'on utilise des données administratives, leur date d'obtention est toujours décalée par rapport à la date de référence. Par exemple, les données fiscales définitives relatives à l'année 2023 arrivent à l'Insee début 2025. En d'autres termes, t3 − t2 n'est pas nul. Dans les faits, les statisticiens n'ont donc de prise que sur le résidu consacré au traitement « statistique » des données, c'est-à-dire t4 − t3.

Commençons alors par nous intéresser au temps de préparation t2 − t1[9], qui a son importance en particulier dans le contexte d'intégration de données nouvelles dans le système d'information statistique.

La phase de préparation

Pour les enquêtes, ménages ou entreprises, concerter pour convaincre

La concertation est la clé de la crédibilité de la statistique publique : en réunissant autour d'une même table producteurs et utilisateurs des statistiques publiques, elle assure qu'un sujet a été correctement évalué dans l'ensemble des dimensions pertinentes. Grâce à elle, le système d'information bâti pour mesurer et analyser un phénomène

9 À noter que t2 − t1 est une approximation du temps de préparation, car t2 est la date de référence des statistiques et non la fin de préparation. C'est donc un minorant, mais c'est un bon ordre de grandeur.

social ou économique résulte bien d'un consensus, dans la limite des moyens alloués. L'objectif est de vérifier l'adhésion aux concepts, ainsi qu'au dispositif qui doit permettre de les mettre en œuvre.

> ***La concertation est la clé de la crédibilité de la statistique publique : en réunissant autour d'une même table producteurs et utilisateurs des statistiques publiques, elle assure qu'un sujet a été correctement évalué dans l'ensemble des dimensions pertinentes.***

On reviendra plus bas sur les instances mises en place dans cet objectif. Cette étape-clé peut s'avérer longue, car elle suppose de la réflexion et une discussion entre personnes venues d'horizons professionnels très différents. C'est souvent un processus itératif, qui demande un certain nombre de va-et-vient entre les statisticiens qui construisent l'outil de mesure et les personnes consultées (***figure 2***).

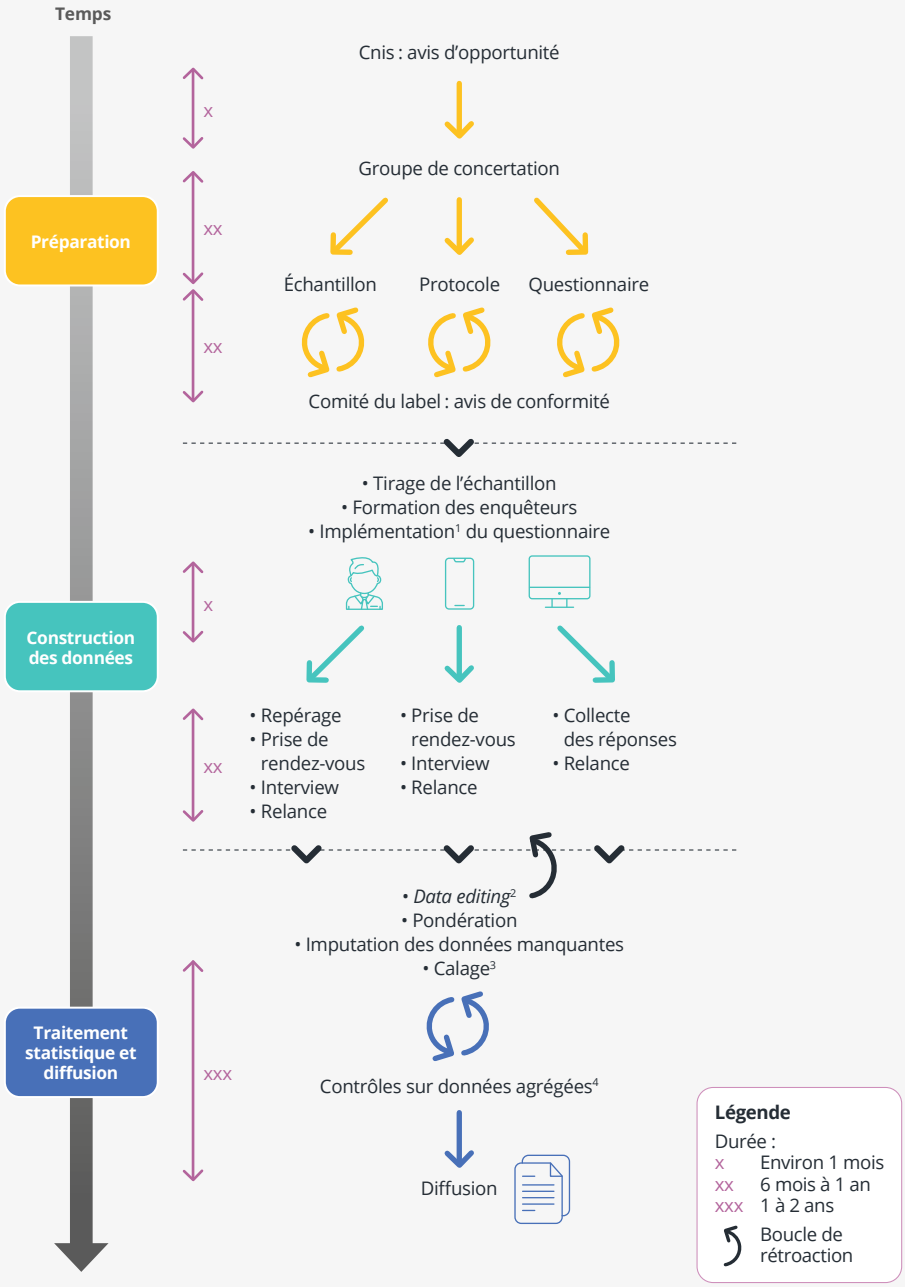
Ainsi, dans le cadre des enquêtes auprès des ménages, l'élaboration du questionnaire résulte d'échanges très poussés entre statisticiens, chercheurs, experts des domaines concernés et partenaires sociaux[10]. Ces considérations s'appliquent naturellement pour des dispositifs d'enquêtes en construction ou en refonte. Mais même pour les enquêtes régulières bien installées, il convient de s'assurer que le questionnaire soumis aux ménages est toujours d'actualité et n'omet pas d'éventuelles nouveautés. Il faut dans bien des cas satisfaire des objectifs parfois contradictoires : construire un questionnaire le plus complet possible, permettant d'analyser l'ensemble des dimensions d'un phénomène donné, et qui soit en même temps intelligible et d'une complexité et d'une longueur maîtrisées. Par exemple, si l'on s'intéresse aux mécanismes de transmission intergénérationnelle du patrimoine, il est important de demander aux personnes répondant à l'enquête Histoire de vie et Patrimoine[11] de décrire leur fratrie et plus particulièrement leur rang de naissance ; ou encore de décrire leurs conditions de vie durant leur enfance et les revenus de leurs parents. Il s'agit donc d'un exercice d'équilibriste, entre richesse de questionnement et attention portée à la facilité de réponse.

Rechercher la pertinence des concepts…

Les concepts sur lesquels reposent les indicateurs statistiques construits à partir de ces données d'enquêtes ne vont pas nécessairement de soi. L'élaboration de ces concepts, dont la nature est largement séculaire, nécessite un effort de réflexion qui peut s'inscrire dans la durée et qui peut naturellement prêter à débat. La pauvreté en est un exemple emblématique : cette notion dépend largement du contexte social et elle varie beaucoup dans l'espace et dans le temps. Il faut alors s'assurer de la pertinence des indicateurs la caractérisant. Les réflexions sur la mesure de la pauvreté débutent dans les années 1970, avec les premiers travaux de chercheurs comme ceux d'Anthony B. Atkinson (Atkinson, 1975). Les recherches se poursuivent durant les années 1980 et 1990. Face au caractère

10 Sans compter les enquêtes encadrées par un règlement de l'Union européenne, qui font également l'objet d'une concertation à ce niveau.
11 Voir le lien : *https://www.insee.fr/fr/metadonnees/source/serie/s1005*.

Figure 2 - La production de statistiques publiques fondées sur une enquête



Cnis : Conseil national de l'information statistique.
1. Programmation (quand il ne s'agit pas d'un questionnaire à remplir sur papier).
2. Processus de vérification des données. Cette étape peut conduire à retourner à la construction des données (rappeler des entreprises...).
3. Processus visant à modifier légèrement les pondérations de manière à retrouver des grandeurs connues (comme le nombre d'individus).
4. Contrôles de cohérence avec les résultats d'autres sources officielles.

multidimensionnel de la pauvreté, plusieurs approches sont développées : pauvreté absolue ou relative, monétaire, en conditions de vie, subjective, administrative… (Insee, 1998). Dans les années 2000, des travaux conséquents ont lieu au niveau européen avec la stratégie de Lisbonne[12]. Ils aboutissent à la mise en place du dispositif d'enquête européen EU-SILC[13]. Ce dernier permet des analyses comparatives selon différentes approches, dont en premier lieu la pauvreté monétaire et la privation matérielle et sociale.

… ainsi que la crédibilité du dispositif

Une fois les concepts bien établis, il faut également concevoir le dispositif de collecte garantissant l'information la plus fiable à moindre coût. Ainsi, une fois que l'on a défini un indicateur de pauvreté basé sur les revenus, il faut trouver la meilleure façon de mesurer ces revenus ; ce que l'on a d'abord fait par le biais d'un questionnaire, dont il a fallu s'assurer qu'il couvrait correctement les différentes composantes du revenu et qu'il était compréhensible du grand public. Ceci passe par le test du questionnaire auprès d'un petit échantillon d'individus ou d'entreprises, qui permet également d'évaluer la bonne compréhension des concepts et la disponibilité de l'information auprès des unités enquêtées.

> ***Pour incorporer les réflexions conceptuelles dans un questionnaire d'enquête, entre les différents tests et les passages devant les instances de concertation, il peut s'écouler jusqu'à un ou deux ans pour un dispositif complètement nouveau.***

Combien de temps alors pour construire un questionnaire permettant d'incorporer les réflexions conceptuelles ? Entre les différents tests et les passages devant les instances de concertation que sont le Conseil national de l'information statistique (Cnis) (Anxionnaz et Maurel, 2021) et le Comité du label (Christine et Roth, 2020), les réflexions peuvent ainsi s'étendre jusqu'à un ou deux ans pour un dispositif complètement nouveau. Le passage devant le Cnis doit permettre d'abord d'examiner l'opportunité du dispositif de collecte dans le contexte social et statistique. Moment clé dans la vie d'une enquête, il permet bien souvent d'initier des premiers contacts avec le monde de la recherche et la société civile, pour constituer ensuite le groupe de concertation qui sera sollicité pour l'élaboration du questionnaire. S'il n'est pas évident de caractériser un questionnaire achevé[14], on peut estimer que la concertation est terminée lorsque l'ensemble des remarques formulées par les parties prenantes, ainsi que les résultats des tests sur le terrain, ont pu être instruits et ont fait l'objet d'arbitrages sur le questionnaire. Décrit ainsi, l'exercice peut sembler simple : il n'en demeure pas moins chronophage et laborieux, l'élaboration d'un questionnaire résultant d'une tension constante entre parcimonie et exhaustivité.

12 Voir le lien : *https://www.europarl.europa.eu/highlights/fr/1001.html*.
13 *European Union statistics on income and living conditions* (statistiques sur les revenus et les conditions de vie dans l'Union européenne). Voir le lien : *https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions*.
14 Dans la mesure où les problématiques qui traversent la société et l'économie sont en évolution constante et que celles-ci peuvent mériter des éclairages qui évoluent au cours du temps.

Données administratives : comprendre l'univers et élaborer un cadre

Lorsque les données sont fournies à la statistique publique, en provenance d'entités administratives, le travail de préparation est radicalement différent : il ne s'agit plus de se mettre d'accord sur des variables à collecter, encore moins sur un questionnaire, puisque le contenu des données est imposé aux statisticiens publics (Hand, 2018). Il faut donc d'abord comprendre un environnement externe, un contenu métier étranger, ce qui requiert un long travail de « découverte » de l'univers en question (***figure 3***). Cette familiarisation avec un autre langage, d'autres manières de travailler, d'autres codes, se révèle nécessaire si l'on veut utiliser ces données sans commettre d'erreur d'interprétation grossière. Elle est à ce point importante que l'*Office for National Statistics* (ONS)[15] est allé jusqu'à mettre en place des détachements de statisticiens, pendant plusieurs mois, au sein d'administrations détentrices des données (Fermor-Dunman et Parsons, 2022).

Cette phase de découverte est l'occasion d'acquérir une bonne connaissance de la source administrative, de son statut, du contexte dans lequel elle s'inscrit (juridique, notamment) et des liens existants avec d'autres entités. De ce point de vue, les discussions peuvent prendre beaucoup de temps. En effet, le cadre juridique dans lequel s'exerce l'activité de la statistique publique n'est pas toujours bien connu des autres administrations. Aussi, la sensibilité des données susceptibles de faire l'objet d'un échange peut conduire leur détenteur à une certaine prudence. Puis, peu à peu, des modalités de coopération s'établissent entre l'administration, en tant que futur « fournisseur » des données, et le service statistique, en tant que futur « récepteur » de ces données.

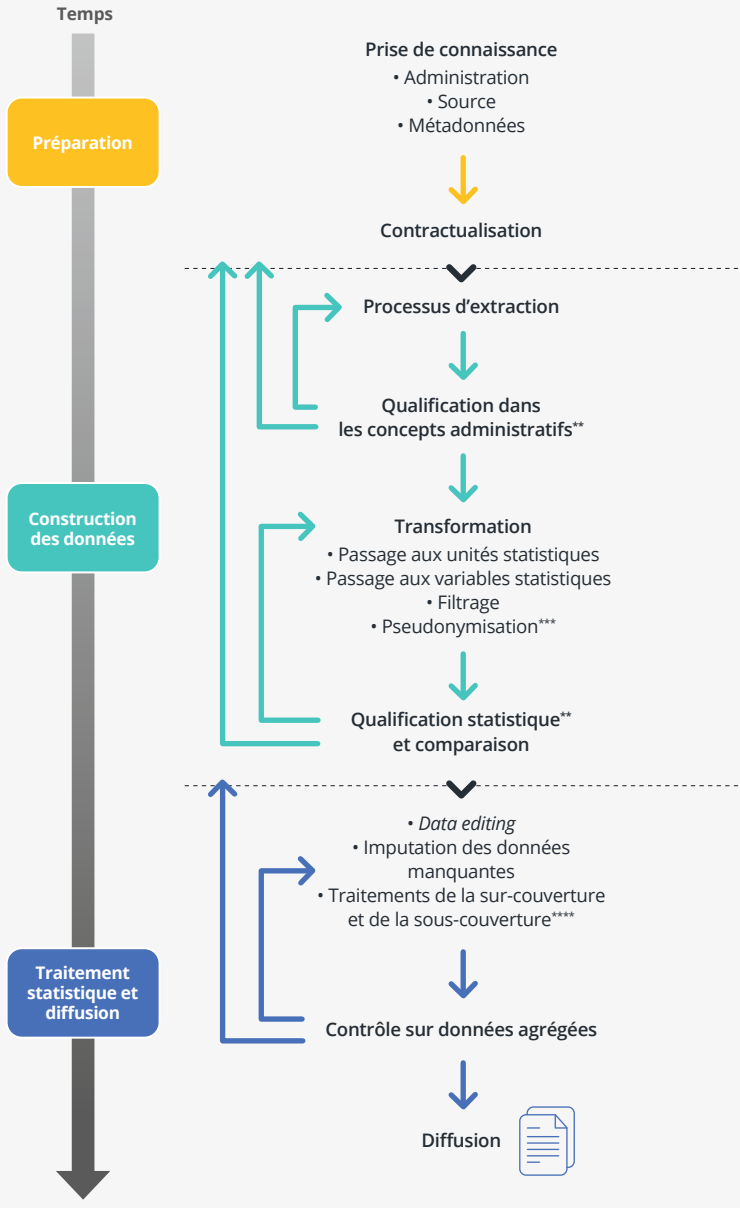
> ***Il ne s'agit pas seulement d'accéder à une documentation « technique », mais également de comprendre ce qui la sous-tend.***

Tout cela passe par une connaissance précise, détaillée, rigoureuse des métadonnées[16] : signification des variables et des nomenclatures utilisées, liste de valeurs prises, caractérisation temporelle, standards utilisés pour les échanges[17]… Cerroni et al. (2014) insistent sur le besoin d'appréhender en premier lieu la source et les métadonnées avant tout travail statistique. Il ne s'agit pas seulement d'accéder à une documentation « technique », mais également de comprendre ce qui la sous-tend. Ainsi, utiliser des données fiscales requiert un minimum de maîtrise de la législation fiscale. Il s'agit aussi de décider des variables nécessaires. Par exemple, une réflexion fine a été menée pour décider des variables issues de fichiers administratifs à intégrer dans le répertoire statistique des individus et des logements (Résil)[18] tout en limitant les risques d'identification des personnes (Lefebvre, 2024).

Le dialogue avec l'administration vise enfin à mettre en place un processus opérationnel de fourniture des données, ce qui n'est pas une mince affaire : jeux de données complexes, volumineux, souvent transmis en plusieurs fois, avec possibilités de retour en arrière si les données sont incomplètes ou erronées. Il en résulte une convention entre administration et service statistique, dont l'élaboration, issue de multiples allers-retours,

15 Institut national de statistique du Royaume-Uni.
16 Une métadonnée est une donnée qui définit et décrit une autre donnée. Voir à ce sujet Bonnans (2019).
17 Par exemple, la norme d'échanges NEODeS pour la déclaration sociale nominative (Humbert-Bottin, 2018).
18 Voir le lien : *https://www.insee.fr/fr/information/7748881*.

▶ Figure 3 - La production de statistiques publiques fondées sur des données administratives – schéma très simplifié*



* Ce schéma est une synthèse simplifiée de trois schémas successifs de Koumarianos et Rivière (2025).
** La première qualification des données, essentielle, veille à s'assurer du bon état des données a priori, à aboutir à une « source », tout en restant strictement dans le cadre des concepts administratifs. La qualification statistique ayant lieu après l'étape de transformation veille quant à elle à s'assurer que les variables et unités statistiques sont conformes et cohérentes avec l'objectif recherché.
*** La pseudonymisation est un traitement de données à caractère personnel de manière qu'on ne puisse pas attribuer les données à une personne physique sans avoir recours à des informations supplémentaires (Wikipedia).
**** Sous-couverture : le champ couvert représente moins que la population cible, « il en manque ».
Sur-couverture : le champ couvert va au-delà de la population cible, « il y en a trop ».

Lecture : Les phases de qualification et de contrôle des données peuvent entraîner des boucles de rétroaction.

peut s'avérer chronophage. Mais on acceptera d'y passer un temps très significatif si l'échange est régulier, pérenne, comme c'est le cas pour la déclaration sociale nominative (DSN) (Renne, 2018).

Données privées : de nouvelles exigences

L'utilisation de données privées peut sembler très similaire à celle des données administratives : dans les deux cas, il s'agit de récupérer une masse d'informations provenant d'un autre environnement, et donc de comprendre un univers métier, d'accéder aux métadonnées, et d'aboutir à la rédaction d'une convention. Tout le temps de découverte, de compréhension de la source, d'obtention des métadonnées, etc., demeure ici valable.

Mais ces sources soulèvent des défis supplémentaires. La fourniture de données entre administrations est d'autant plus naturelle que celles-ci sont de plus en plus soumises à une exigence d'ouverture. Ce n'est pas le cas des acteurs privés, qui sont dans les faits beaucoup moins tenus de transmettre des informations, et qui n'en voient pas nécessairement l'utilité. La donnée est par ailleurs bien identifiée par ces acteurs privés comme un actif de plus en plus stratégique. Il n'est donc pas simple de les convaincre de partager cet actif avec une administration, sans intérêt évident.

> ***Ici, le dialogue relève donc aussi, en quelque sorte, de la négociation, dans laquelle il faut penser aux contreparties.***

Ici, le dialogue relève donc aussi, en quelque sorte, de la négociation, dans laquelle il faut penser aux contreparties. En particulier, accepte-t-on ou non de payer les données, comme le fait par exemple l'*Instituto Nacional de Estadística* (INE)[19] auprès d'opérateurs espagnols de téléphonie mobile ? Si oui, quels sont les critères pour déterminer dans quels cas on paye ? Si non, il faut que l'entreprise y trouve un autre intérêt, et d'autres contreparties sont alors possibles (Joubert, 2025). Au niveau européen, un effort a également été entrepris pour accéder à certaines données existantes au niveau supranational, comme en témoigne le projet sur les plateformes d'économie collaborative d'Eurostat (Eurostat, 2025).

Quel que soit l'accord mis en place, comme pour les sources administratives, le statisticien doit comprendre le contexte de génération de la donnée, ce qui nécessite de nombreux échanges avec le « propriétaire ». Ceux-ci se feront de manière plus naturelle dans un cadre de coopération, à travers un partenariat entre statistique publique et entreprise privée détentrice des données. L'accord doit par ailleurs intégrer des exigences de confidentialité particulière : à celle des données relatives aux personnes, s'ajoute ici le secret des affaires, qui limite, ou plutôt encadre, l'acceptation du propriétaire à transférer des données (Bonnet et Loisel, 2024).

La convention, cadre contractuel contraignant, est donc tout à fait essentielle. Mais elle sera beaucoup plus longue à réaliser et à mettre en œuvre, comme ce fut le cas pour les données de caisse (Leclair, 2019). Souvent, on passe par des phases d'expérimentation,

19 Institut national de statistique de l'Espagne.

nécessitant de faire intervenir d'autres partenaires, tels que des instituts de recherche (Boittelle et al., 2025) : la multiplicité des acteurs allonge d'autant plus le processus. Par ailleurs, rien ne garantit que la convention soit pérenne : par exemple, la fourniture gratuite de données de téléphonie mobile par Orange a été possible en 2020, au plus fort de la crise sanitaire (Tavernier, 2020), mais n'a pas été poursuivie.

Ainsi, quel que soit le dispositif, il existe toujours un temps significatif de préparation, de cadrage, de négociation : un investissement indispensable avant tout travail effectif sur les données. Pour des opérations pérennes, régulières, celui-ci bénéficie à toutes les opérations ultérieures, et les travaux de préparation sont appelés à se répéter dans le temps, avec une ampleur généralement plus faible qu'au départ.

▶ La phase de construction des données

On entre ici dans une phase plus opérationnelle, le temps t3 − t2, où l'on va s'attacher à construire les données individuelles : soit à travers un processus de collecte (enquêtes auprès des ménages ou des entreprises), soit via un travail d'analyse, de qualification et de transformation de données obtenues par ailleurs (sources administratives ou privées).

Enquêtes auprès des ménages : un temps de collecte incontournable

Pour ce qui est des enquêtes, la statistique publique a une maîtrise assez complète de ce processus, qu'elle passe par un prestataire pour réaliser le terrain ou qu'elle se charge de celui-ci par ses moyens propres. Néanmoins, ce temps n'est pas le même pour toutes les opérations. En effet, mesurer un phénomène saisonnier comme la consommation nécessite une collecte sur une année complète. Par ailleurs, mesurer un phénomène structurel – objet d'enquêtes souvent très espacées dans le temps, voire ponctuelles – suppose de collecter plus d'information. En effet, le processus de concertation porte en général sur un ensemble de variables plus large. Pour ces enquêtes, il faut également davantage de préparation du terrain dans la mesure où le réseau d'enquêteurs n'est pas toujours familier du questionnaire.

De manière générale, la collecte peut se découper en différentes phases. Dans un premier temps, il y a la **phase d'organisation**, qui peut être plus ou moins courte selon que l'enquête est un dispositif récurrent ou non. Elle inclut l'élaboration du protocole de collecte et la rédaction d'une documentation le décrivant. Cette dernière englobe notamment les consignes aux enquêteurs. Il faut également former l'ensemble du réseau des enquêteurs, ce qui peut prendre du temps. Ainsi, lorsque le réseau de l'Insee est sollicité, il faut d'abord former les gestionnaires au niveau national, lesquels forment à leur tour les enquêteurs dans chaque direction régionale. Le processus de formation peut donc prendre un peu de temps, un mois, voire plus dans le cas d'enquêtes comme Sans Domicile[20] : à la fois parce qu'il se fait de manière décentralisée et parce qu'il faut accommoder les agendas très chargés du réseau d'enquêteurs.

20 Voir le lien : https://www.insee.fr/fr/metadonnees/source/serie/s1002.

L'organisation inclut également des étapes essentielles comme la sélection de l'échantillon de personnes à enquêter, l'impression des fiches-adresses (coordonnées des personnes échantillonnées) et leur répartition entre les enquêteurs. L'envoi des lettres-avis annonçant l'enquête auprès des ménages est aussi un jalon essentiel, qui doit s'anticiper pour tenir compte des délais d'acheminement du courrier. Si ces éléments préparatoires peuvent, pour certains, se dérouler en parallèle, la programmation infra-annuelle des enquêtes impose un calendrier permettant d'identifier bien en amont du terrain les zones d'enquête concernées et la volumétrie associée.

> ***Dans une collecte en face-à-face ou par téléphone, les enquêteurs doivent entrer en contact avec les ménages sélectionnés, les convaincre de répondre à l'enquête et souvent prendre rendez-vous avec eux.***

Puis vient la **phase du terrain** proprement dite, qui dépend fortement du mode de collecte. Dans une collecte en face-à-face ou par téléphone, les enquêteurs doivent entrer en contact avec les ménages sélectionnés, les convaincre de répondre à l'enquête et souvent prendre rendez-vous avec eux. Lorsque la taille du réseau d'enquêteurs est fixe, le temps associé à cette phase dépend directement de la volumétrie totale de l'échantillon. Par ailleurs, le protocole peut être plus ou moins chronophage selon le mode de collecte. Ainsi, une collecte en face-à-face nécessite une étape préliminaire, dite de repérage, pendant laquelle l'enquêteur s'assure sur le terrain de l'existence du logement et s'attache à entrer en contact avec ses occupants. Aujourd'hui, les données de contact sont généralement disponibles dans les bases de sondage utilisées pour les enquêtes « ménages », ce qui rend cette phase moins cruciale. Elle reste cependant importante, car elle demeure un moyen essentiel pour atteindre les ménages concernés par l'enquête.

De manière générale, la phase de prise de contact peut durer, car les ménages ne sont pas toujours faciles à joindre. Notamment, certains profils peuvent être rarement présents à leur domicile durant les temps de collecte. Le caractère représentatif de l'enquête, et donc sa qualité, sont donc déterminés par le temps et l'effort passés à contacter l'ensemble des ménages de l'échantillon, qu'ils soient faciles ou non à contacter. Dans le cadre d'une collecte par web, ces questions se posent moins. Mais la collecte uniquement selon ce mode n'est pas un protocole envisageable pour des raisons de représentativité. En effet, les taux de réponse pour la collecte web sont très largement inférieurs à ceux d'une collecte intermédiée par un enquêteur, le rôle de ce dernier étant déterminant dans le processus d'adhésion à l'enquête. Le web est plutôt mobilisé dans le cadre de protocoles multimodes (Beck et al., 2022). Dans le cas où ces derniers sont mobilisés de manière séquentielle, par exemple web, sinon téléphone, il est nécessaire de finir la collecte dans un mode avant de passer à un autre, ce qui aboutit à rallonger significativement le temps de collecte.

Enfin, de façon très variable selon les enquêtes, il existe un **travail d'apurement réalisé par les gestionnaires** : vérification de la bonne compréhension des consignes par les enquêteurs, ainsi que de la cohérence de quelques données collectées. Cette phase, si elle se déroule pour une bonne partie durant la collecte proprement dite, peut nécessiter des allers-retours entre les gestionnaires et les enquêteurs, et dans certains cas avec les enquêtés (Forteza et García-Uribe, 2025).

Enquêtes auprès des entreprises : la place centrale du *data editing*

En matière de collecte, la réalisation d'enquêtes auprès d'entreprises présente des différences importantes par rapport à celles effectuées auprès de ménages. Tout d'abord, pour l'essentiel, ce sont des questionnaires auto-administrés[21], historiquement transmis par courrier, et désormais quasi systématiquement sur le web. Le temps de l'organisation du réseau d'enquêteurs et des interactions entre enquêteurs et enquêtés ne joue donc plus. Il y a bien un suivi d'avancement de la collecte, mais d'une autre nature : on peut connaître à tout moment la liste des non-répondants, ce qui permet d'organiser auprès d'eux un rappel, éventuellement en plusieurs étapes. Cependant, si les relances améliorent le taux de réponse et donc la précision, elles détériorent le délai : c'est donc un compromis à trouver.

> ***Ce qui caractérise les enquêtes auprès d'entreprises, c'est l'importance considérable du* data editing*, c'est-à-dire du processus de vérification des données, automatique ou manuel, et des modifications que cela induit.***

Ce qui caractérise également les enquêtes auprès d'entreprises, c'est l'importance considérable du *data editing*, c'est-à-dire du processus de vérification des données, automatique ou manuel, et des modifications que cela induit. Sur ce sujet, toute une littérature s'est développée depuis longtemps dans la communauté statistique internationale (Granquist, 1997 ; De Waal et al., 2011 ; Pannekoek et al., 2013)[22]. Dans le cas des enquêtes « entreprises », il s'agit de vérifier non seulement les questionnaires jugés douteux, mais aussi les données à fort impact sur les agrégats ; car l'univers observé est très hétérogène, et une seule erreur peut tout dégrader (Lawrence et McKenzie, 2000).

Contrairement aux enquêtes « ménages » où l'on peut souvent imaginer d'automatiser le contrôle et l'imputation[23], opérations qui relèvent alors plutôt des traitements statistiques, le contrôle manuel se révèle ici indispensable et prend inévitablement du temps. À nouveau, c'est un compromis à déterminer entre précision statistique et délai de production (Granquist et Kovar, 1997). Ce temps de vérification porte sur toutes les « cases » des tableaux finaux. En effet, la demande des utilisateurs, par exemple celle des fédérations professionnelles, peut concerner un secteur d'activité très précis (au niveau le plus détaillé de la nomenclature d'activités française (NAF)[24], par exemple). Dès lors, le fait que les statistiques produites portent sur des granularités[25] très fines conduit mécaniquement à un temps de travail significatif pour les gestionnaires d'enquête, car proportionnel au nombre de cases de diffusion. Il est cependant possible de prioriser le travail pour corriger avant tout les erreurs les plus influentes sur les résultats, grâce à

21 Remplis directement par les enquêtés, sans l'intermédiation d'un enquêteur.
22 Caron (2025) donne une première approche des principes généraux de mise en œuvre du *data editing* qui existent dans la littérature.
23 Voir l'article fondateur de Fellegi et Holt (1976). L'imputation est une technique statistique utilisée pour remplacer les données manquantes par des valeurs de substitution, permettant ainsi d'obtenir un ensemble de données plus complet pour l'analyse. L'exemple de Forteza et García-Uribe (2025) cité plus haut montre cependant l'intérêt d'un traitement manuel dans les enquêtes « ménages » pour des variables dont la distribution est très asymétrique.
24 Voir le lien : *https://www.insee.fr/fr/information/2406147*.
25 Pour les statistiques, la granularité correspond à la finesse décrite par les « cases » des tableaux : par région, département, commune ? Par tranche d'âge de 5 ans, 10 ans ? Par secteur d'activité en 21 sections, 88 divisions, ou 615 classes ?

des procédures de vérification sélective des données (*selective editing*). Ainsi, les équipes en charge du dispositif Ésane[26] ont mis en œuvre une telle méthode, à l'appui du calcul de scores permettant de caractériser la criticité des variables à contrôler.

En pratique, pour les variables à contrôler in fine, le travail des gestionnaires d'enquête va souvent consister à rappeler les unités enquêtées, afin de compléter ou corriger la réponse initiale, ou de mieux comprendre les données transmises et s'assurer qu'il n'y a pas d'erreur. Ce retour à l'enquêté n'existe pas pour les ménages. D'abord parce que l'interaction avec l'enquêté s'effectue à la source, lors du questionnement. Mais aussi parce que l'univers des ménages est beaucoup plus homogène, de sorte qu'une erreur est moins préjudiciable pour la qualité des résultats de l'enquête.

Données administratives : itérations avec l'administration et transformations

Il peut paraître étrange de parler de temps passé à la « collecte » dans le cas des données administratives. En effet, on pourrait considérer que, la convention ayant été écrite, il suffit de l'appliquer pour récupérer les données, avant de passer à la tabulation proprement dite. Le temps nécessaire serait donc, sinon nul, du moins négligeable. Ce serait méconnaître la réalité !

> ***Les données administratives ont une finalité bien précise. Elles servent avant tout à faire fonctionner des processus opérationnels, à piloter, ou à prendre des décisions. Elles possèdent donc une structure liée à ces objectifs.***

Les données administratives ont une finalité bien précise. Elles servent avant tout à faire fonctionner des processus opérationnels, à piloter, ou à prendre des décisions. Elles possèdent donc une structure liée à ces objectifs. Le traitement qu'il faut leur faire subir pour les intégrer dans un processus statistique est très généralement conséquent, et la conception de celui-ci prend du temps. Il faut bien souvent un certain nombre d'allers-retours entre les administrations détentrices de ces données et la statistique publique avant d'arriver à identifier les données pertinentes, puis les transformations qu'il faut leur imposer… afin de convertir une donnée à visée opérationnelle en une donnée statistique. Cela passe aussi par une phase de consolidation de la donnée visant à éliminer les doublons, lesquels ne manquent pas d'apparaître dans des données à visée opérationnelle.

Par ailleurs, du côté de l'administration, l'**élaboration de la donnée administrative** peut se révéler assez longue, et c'est là un délai de disponibilité sur lequel la statistique publique n'a aucune prise. Ainsi, les données fiscales sur les revenus des ménages ne sont disponibles qu'un an après la fin de la période de référence. En effet, pour mesurer le revenu des ménages dans l'année N, il faut d'abord que ces derniers réalisent leur déclaration de revenus au printemps N+1. Vient ensuite le traitement de l'information par l'administration fiscale, qui inclut les correctifs liés aux déclarations tardives et aux contrôles. Celui-ci dure jusqu'à la fin de l'année N+1.

26 Élaboration des statistiques annuelles d'entreprises. Voir le lien : *https://www.insee.fr/fr/metadonnees/source/serie/s1188*.

De ce fait, la statistique publique dépend du processus de recueil de la donnée administrative. Elle peut parfois décider de recourir à des données plus provisoires, au prix d'une qualité moindre de l'information. C'est par exemple ce qui a été fait pour les données sur les revenus intégrées dans le dispositif Statistiques sur les revenus et les coûts (SRVC)[27], partie française de EU-SILC. En effet, afin d'atteindre les nouvelles exigences en matière de fraîcheur demandées par Eurostat, il a été décidé de recourir à une information fiscale moins consolidée, de manière à gagner quelques mois dans le processus d'élaboration de la donnée statistique.

Néanmoins, l'essentiel de l'effort va porter sur les travaux méthodologiques nécessaires à la construction du processus de **transformation de la donnée opérationnelle en donnée statistique** (Koumarianos et Rivière, 2025). L'objectif est d'adapter l'information collectée par l'administration aux concepts et au cadre de référence de la statistique publique. Ainsi, dans le Fichier démographique sur les logements et les individus (Fidéli)[28] (Lamarche et Lollivier, 2021), l'identification des résidences principales fait appel à des concepts issus du recensement de population, définis par l'Insee (et plus ou moins facilement mis en œuvre dans les processus de collecte dont il a la maîtrise). Mais la source fiscale, elle, s'appuie sur une notion de résidence principale très adhérente à la problématique du recouvrement de l'impôt. Toute la difficulté réside donc dans la capacité à adapter l'information sur la localisation des individus dans la source fiscale pour se rapprocher le plus possible du concept statistique.

Données privées : un cadre plus contraignant pour la transformation des données

Pour les données privées, la situation est très similaire à celle des données administratives. Ce qui change, c'est l'exigence accrue de confidentialité des données. Le statisticien public est souvent conduit à travailler directement sur un poste de travail de l'entreprise concernée, car celle-ci n'envoie pas de fichier de données individuelles, même anonymisées. C'est le cas par exemple pour les données de comptes bancaires (Bonnet et Loisel, 2024). Les entreprises peuvent aussi souhaiter que les statisticiens n'accèdent pas directement aux données des personnes, mais uniquement à des données transformées, comme pour les données de téléphonie mobile (Joubert, 2025). Le développement d'un *pipeline* conçu en collaboration entre l'entreprise et les statisticiens publics – pour s'entendre sur la méthode de transformation – est alors particulièrement crucial (Lesur, 2025).

> ***Nouvel usage implique nouvelle convention.***

Le respect rigoureux des conventions constitue également une contrainte forte : on ne peut utiliser les microdonnées[29] hors du cadre invoqué dans la convention simplement parce qu'on considère que c'est intéressant. Nouvel usage implique nouvelle convention. Ce fut le cas, par exemple, pour les données de caisse : si ces dernières contribuent de manière réglementaire au calcul de l'indice des prix à

27 Voir le lien : *https://www.insee.fr/fr/metadonnees/source/serie/s1220*.
28 Voir le lien : *https://www.insee.fr/fr/metadonnees/source/serie/s1019*.
29 Données individuelles, non agrégées : par exemple, des données par ménage, et non par catégorie de ménage, ou des données par entreprise, et non par secteur d'activité.

la consommation (Leclair 2019), elles peuvent être utiles également pour estimer la consommation des ménages, ou encore le chiffre d'affaires des grandes enseignes de distribution.

▶ La phase finale de traitement statistique

Corriger les processus de sélection, en particulier dans les enquêtes

Pour les enquêtes, à l'issue de l'étape de collecte, on dispose de microdonnées « originelles », qu'on qualifie souvent de « brutes »[30], et qui ne sont pas les microdonnées finales. En particulier, les données présentent de manière très générale un certain nombre de manques, qu'il va falloir redresser de manière statistique.

> ***Pour les enquêtes, à l'issue de l'étape de collecte, les données présentent de manière très générale un certain nombre de manques, qu'il va falloir redresser de manière statistique.***

Le premier d'entre eux vient du fait que les enquêtes portent le plus souvent sur des **échantillons**, qui doivent donc être pondérés pour être « représentatifs » de la population dans laquelle ils sont sélectionnés (au sens de refléter au final sa structure et ses effectifs).

Cependant, à la sélectivité maîtrisée par le statisticien qu'est l'échantillonnage, s'ajoutent d'autres phénomènes de sélection, tels que la **non-réponse totale**. Le fait qu'un ménage ou une entreprise refuse de répondre, en dépit de l'obligation légale, est une réalité avec laquelle la profession statistique doit composer. Il faut alors comprendre le mécanisme de non-réponse – les déterminants observables de l'absence de réponse – et incorporer ces déterminants dans le calcul des pondérations. Cette étape est essentielle pour préserver la « représentativité » de l'échantillon de répondants (au sens de la capacité à reconstruire une image fidèle de la population). La compréhension du mécanisme de « non-réponse » ne peut pas se faire a priori : dans bien des situations, ce phénomène reste difficile à anticiper et nécessite une expertise statistique.

Remplir les « trous », une phase potentiellement fastidieuse

Il faut ensuite faire de même pour ce qu'on appelle la **non-réponse partielle**. Un ménage ou une entreprise échantillonnée – même en ayant accepté de répondre – peut pour une raison ou une autre ne pas fournir de réponse à une question particulière : par choix, parce que la personne interrogée ne sait pas, ou parce que la question ne la concerne pas. Ce phénomène est généralement traité par imputation. On va chercher à comprendre les déterminants de la non-réponse, en étudiant la corrélation entre la variable non collectée et les autres. Il s'agit ensuite d'attribuer la valeur la plus probable au regard des informations connues pour l'unité enquêtée. Le travail peut être très important, non

30 Même si le terme est inadéquat, car les données sont toujours une construction, et ne sont en réalité jamais « brutes » (Gitelman, 2013).

seulement en matière de modélisation des mécanismes de non-réponse, mais aussi selon l'exigence d'exhaustivité de la correction. En effet, l'ensemble des variables collectées peut être affecté par la non-réponse partielle. Il existe des méthodes bayésiennes[31] pour redresser de manière systématique tout un corpus de variables (Rubin, 1987), mais celles-ci n'exonèrent pas d'une analyse souvent fastidieuse des mécanismes de non-réponse affectant chacune des variables.

La situation est très similaire pour les enquêtes et pour les données administratives ou privées. Simplement, les « données de base » ne sont pas issues d'une collecte, mais, comme on l'a vu, d'un processus de récupération de données du système d'information, et de transformations. Néanmoins, les « traitements statistiques » sont les mêmes.

Le *fine tuning*[32] des pondérations des données

Reste enfin, pour les enquêtes, le « calage sur marges », qui est un processus permettant de modifier légèrement les pondérations de manière à retrouver dans l'échantillon de répondants des grandeurs connues (comme le nombre d'individus). Si la loi des grands nombres et la théorie du sondage assurent que ces grandeurs seront approchées de manière satisfaisante, il peut être efficace de caler l'échantillon sur telle ou telle grandeur, de manière à améliorer la précision de l'échantillon sur les variables d'intérêt de l'enquête. Ce processus de calage peut lui aussi être fastidieux et impliquer un nombre élevé de variables. Pour les statistiques d'entreprises, se posent également des problèmes spécifiques liés aux unités à fort impact sur les agrégats publiés, nécessitant un travail approfondi sur les pondérations.

À première vue, l'ensemble de ces traitements (contrôle, imputation, repondération) se ramène finalement à une transformation. Dans l'idéal, pour gagner du temps, on pourrait imaginer de l'automatiser, c'est-à-dire de développer un *pipeline*. Une fois conçu et mis au point, celui-ci n'aurait plus qu'à être appliqué de manière systématique, comme tout traitement de données mis au point par un *data scientist* (Comte et al., 2022). Mais de tels travaux exigent une analyse qui peut s'avérer chronophage, selon le niveau d'exigence en matière de complétude et de cohérence. Ils peuvent ainsi requérir un temps significatif lorsqu'ils sont mis en œuvre. Il n'en demeure pas moins qu'un effort en matière de standardisation des outils et des pratiques est déjà un gage d'optimisation.

Exigences de cohérence et limites de l'automatisation

Attardons-nous sur les exigences de cohérence qui pèsent sur les données produites par la statistique publique, afin de mieux comprendre les obstacles à une automatisation complète.

On pense d'abord aux « **microcontrôles** », dans lesquels on s'assure que les données relatives à une même unité statistique ne présentent pas d'incohérences manifestes entre elles. On a déjà évoqué, dans le cas des enquêtes auprès des entreprises, l'enjeu

31 Méthodes s'appuyant sur le théorème de Bayes, qui permettent d'estimer la probabilité de phénomènes à partir de la connaissance de certaines informations.
32 Réglage fin.

> ***Les exigences de cohérence qui pèsent sur les données produites par la statistique publique rendent difficiles une automatisation complète des traitements.***

d'un traitement manuel de certaines de ces incohérences, en particulier quand l'impact de la réponse de l'entreprise sur les résultats agrégés peut être sensible.

Mais on ne peut omettre les **contrôles sur données agrégées**, dans lesquels on met en regard les statistiques obtenues avec d'autres informations : statistiques des périodes précédentes, relatives à d'autres pays ou régions, ou bien provenant d'autres sources officielles. Cette cohérence est un critère essentiel pour juger de la qualité des statistiques et pour les qualifier également d'officielles. C'est d'ailleurs le quatorzième principe du Code de bonnes pratiques de la statistique européenne. Toute source d'incohérence peut être une raison légitime de discrédit des chiffres produits.

Or, les exigences de cohérence peuvent porter sur des domaines de diffusion[33] qui peuvent être très fins. En effet, non seulement la statistique publique publie ses propres analyses, mais elle met également les microdonnées à disposition du monde de la recherche. Ainsi, les statisticiens doivent s'assurer en amont, autant que possible, que les données produites peuvent être utilisées sans contredire de manière rédhibitoire d'autres statistiques publiées – ou à tout le moins comprendre et fournir des éléments d'explication aux contradictions apparentes.

Par exemple, les sources statistiques contenant de l'information sur les revenus sont multiples. En effet, pour de nombreuses enquêtes « ménages », il s'agit d'une dimension d'analyse très importante, soit parce qu'elle est au cœur des thématiques de l'enquête, soit parce que c'est une source d'explication déterminante pour les phénomènes auxquels elle s'intéresse. Il faut alors s'assurer que les chiffrages implicitement fournis par chacune de ces enquêtes sont cohérents avec le dispositif de référence sur le revenu[34].

> ***Même avec un* pipeline *automatisant les contrôles de cohérence, la majorité du temps passé consistera à analyser les résultats, à comprendre d'où viennent les éventuelles incohérences et à déterminer sur quoi intervenir.***

De la même manière, il faut s'assurer d'une certaine cohérence entre information macroéconomique (issue des comptes nationaux) et information microéconomique (issue des enquêtes). Cette exigence s'est renforcée ces quinze dernières années, à la suite du rapport de la commission Stiglitz-Sen-Fitoussi sur la mesure de la performance économique et du progrès social (Stiglitz et al., 2009). Ce rapport insiste en effet sur la nécessité, pour la statistique publique, de mesurer à la fois les agrégats économiques et leur distribution au sein de la population.

33 Regroupements au niveau desquels les données sont agrégées pour être diffusées, i.e. les « cases » de tableaux évoquées plus haut : par exemple, regroupement selon le croisement département / tranche d'âge.
34 Qui dans ce cas précis est l'enquête Revenus Fiscaux et Sociaux (ERFS).

Ainsi, même si on a réalisé un *pipeline* automatisant les contrôles de cohérence, la majorité du temps passé consistera à en analyser les résultats, à essayer de comprendre d'où proviennent les éventuelles incohérences et à déterminer sur quoi intervenir : enlever une unité douteuse ou hors champ, compléter le champ, revoir des modalités de codage, etc. C'est là un travail de fourmi, et chercher à trop réduire cette durée ferait courir des risques concernant la qualité.

Quels enseignements tirer ?

L'explication du temps nécessaire à l'élaboration de statistiques de qualité, vue comme une fonction de plusieurs variables explicatives, est paradoxalement difficile à fonder sur des données (i.e. difficile à modéliser). Comme le souligne Parkinson (1955), « le travail s'étale de façon à occuper le temps disponible pour son achèvement »[35]. En d'autres termes, si on nous octroie plus de temps pour une tâche, on peut avoir tendance à remplir le temps en question. Une mesure fondée sur des temps observés est donc sujette à caution. C'est la raison pour laquelle le présent article tente plutôt de raisonner de façon qualitative ou par ordre de grandeur.

Dès lors, quels sont les facteurs explicatifs de ce temps (***figure 4***) ?

- Lors de la **préparation**, ce sont la complexité du sujet, son caractère polysémique, mais aussi le manque de standardisation des données, la sensibilité des données et les délais de livraison propres au fournisseur, en cas de sources externes. Pour une

Figure 4 - Facteurs influant sur le temps de production des statistiques publiques



	ENQUÊTES	SOURCES ADMINISTRATIVES	SOURCES PRIVÉES
PRÉPARATION	• Nombre de variables • Caractère conjoncturel ou structurel de l'enquête • Complexité du sujet	• Nombre de variables • Sensibilité des données • Standardisation de la donnée*	• Nombre de variables • Sensibilité et intérêt stratégique des données • Standardisation de la donnée* • Champ
CONSTRUCTION DES DONNÉES	Nombre d'unités enquêtées (selon le mode de collecte)	Volatilité de la donnée	Volatilité de la donnée
TRAITEMENT STATISTIQUE ET DIFFUSION	• Nombre de variables • Exigence de cohérence • Granularité de diffusion**	• Nombre de variables • Exigence de cohérence • Granularité de diffusion**	• Nombre de variables • Exigence de cohérence • Granularité de diffusion**

* Transformation pour adapter les données aux concepts et au cadre de référence de la statistique publique.
** Niveau de détail le plus fin des données diffusées.

35 L'analyse de l'auteur est connue sous le nom de « loi de Parkinson ».

opération régulière, dès la seconde opération, on va clairement gagner du temps de compréhension des données, sauf si la réglementation évolue significativement. Les données administratives ou privées peuvent s'avérer très chronophages car il s'agit de s'assurer, avant transformation, que l'on comprend bien le sens et les limites des données.

• **La construction des données**, lorsqu'il s'agit d'enquêtes, est sans doute la phase la plus maîtrisable dans un institut statistique, où les procédures de collecte sont très organisées. Le nombre d'unités statistiques est alors décisif, ainsi que la complexité du questionnement. Cela dépend du protocole : pour les enquêtes par Internet, donc sans l'intermédiation d'un enquêteur, ce temps de collecte sera court. Cependant, on « paie » ce gain de temps ultérieurement, par une étape méthodologique supplémentaire : prise en compte de la baisse du taux de réponse et du biais de sélection, harmonisation avec les résultats issus d'autres modes de collecte. Car pour produire des statistiques publiques, il est aujourd'hui difficile d'imaginer collecter de l'information exclusivement par Internet. Pour la construction de données issues de sources externes, un *pipeline* de traitement automatique, réplicable, et un environnement facilitant sa mise au point améliorent les choses. Mais dans tous les cas, enquêtes ou sources externes, il faut déjà consacrer des ressources aux vérifications : par les enquêteurs, par les gestionnaires et par les statisticiens en charge du dispositif, par exemple, s'il faut rappeler des entreprises parce que leurs réponses paraissent incohérentes.

• **Lors du traitement statistique**, on pourra certes réaliser un autre *pipeline*, menant cette fois des microdonnées aux statistiques, qui peut permettre de gagner un peu de temps par la systématisation des opérations et par la mutualisation plus ou moins forte des outils. Mais il sera inévitablement interrompu par des interventions manuelles, car on devra en permanence s'assurer de la cohérence des résultats. Pour chaque variable, pour chaque domaine de diffusion, on devra regarder s'il y a une rupture de série ou s'il existe des écarts significatifs avec d'autres statistiques et, si oui, chercher à comprendre pour quelle raison. La statistique n'est pas produite seule, elle s'inscrit dans un environnement, c'est une pièce dans un puzzle. Il faut pouvoir interpréter, donner un sens, et plus on a de variables et de strates de diffusion, plus longues et nombreuses sont les vérifications.

Avec la prégnance de l'activité de vérification, et le temps passé dans les boucles de rétroaction (*feedback loops*) qui lui sont associées, il paraît illusoire de gagner beaucoup de temps « à exigences constantes ». En jouant sur les exigences, en revanche, « si l'enjeu du temps est particulièrement important », on peut trouver un compromis : en réduisant le nombre de variables, de questions, ou bien en produisant rapidement des statistiques sur un petit nombre d'entre elles. On peut aussi imposer des limites sur les usages à granularité de diffusion fine. D'une façon ou d'une autre, gagner du temps, c'est accepter de « perdre » sur l'une ou l'autre des dimensions…

> ***Avec la prégnance de l'activité de vérification, et le temps associé, il paraît illusoire de gagner beaucoup de temps « à exigences constantes ».***

▶ Bibliographie

- ANXIONNAZ, Isabelle et MAUREL, Françoise, 2021. Le Conseil national de l'information statistique – La qualité des statistiques passe aussi par la concertation. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 123-142. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5398693?sommaire=5398695*.

- ATKINSON, Anthony B., 1975. *The Economics of Inequality*. Mai 1975. Clarendon Press, Oxford University Press. ISBN 0-19-877076-6.

- BECK, François, CASTELL, Laura, LEGLEYE, Stéphane et SCHREIBER, Amandine, 2022. Le multimode dans les enquêtes auprès des ménages : une collecte modernisée, un processus complexifié. In : *Courrier des statistiques*. [en ligne]. 20 janvier 2022. Insee. N° N7, pp. 7-28. [Consulté le 6 octobre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6035934?sommaire=6035950*.

- BÉNICHOU, Yves-Laurent, ESPINASSE, Lionel et GILLES, Séverine, 2023. Le code statistique non signifiant (CSNS) : un service pour faciliter les appariements de fichiers. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp. 64-85. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635825?sommaire=7635842*.

- BOITTELLE, Mathieu, CUPILLARD, Émilie, JACQUOT, Alain, JOUBERT, Marie-Pierre et LE GOFF Florian, 2025. Les données de transaction par carte bancaire CB – Quels apports possibles aux analyses conjoncturelles et territoriales ? In : *Courrier des statistiques*. [en ligne]. 23 juin 2025. Insee. N° N13, pp. 119-142. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8546947?sommaire=8546949*.

- BONNANS, Dominique, 2019. RMéS, le référentiel de métadonnées statistiques de l'Insee. In : *Courrier des statistiques*. [en ligne]. 27 juin 2019. Insee. N° N2, pp. 46-57. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4168396?sommaire=4168411*.

- 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 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8264558?sommaire=8264562*.

- CARON, Nathalie, 2025. Le *data editing* : définition et principes généraux. In : *Documents de travail*. [en ligne]. 20 octobre 2025. Insee. N° M2025/06. [Consulté le 3 novembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/8658331*.

- CERRONI, Fulvia, DI BELLA, Grazia et GALIÈ, Lorena, 2014. Evaluating administrative data quality as input of the statistical production process. In : *Rivista di statistica ufficiale*. [en ligne]. Octobre 2014. Istat. Volume 16, N° 1-2, pp. 117-146. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.istat.it/produzione-editoriale/rivista-di-statistica-ufficiale-1-22014/*.

- CHRISTINE, Marc et ROTH, Nicole, 2020. Le Comité du label – Un acteur de la gouvernance au service de la qualité des statistiques publiques. In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. Insee. N° N5, pp. 39-52. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5008698?sommaire=5008710*.

- 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. 68-85. [Consulté le 15 septembre 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 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635829?sommaire=7635842*.

- De WAAL, Ton, PANNEKOEK, Jeroen et SCHOLTUS, Sander, 2011. *Handbook of Statistical Data Editing and Imputation*. Janvier 2011. John Wiley & Sons, Inc.. ISBN 9780470542804.

- ELBAUM, Mireille, 2018. Les enjeux des nouvelles sources de données. In : *Chroniques du Cnis*. [en ligne]. Septembre 2018. N° 16. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.cnis.fr/publications/liste-des-chroniques-du-cnis/*.

- EUROSTAT, 2025. Short-stay accommodation offered via online collaborative economy platforms. In : *site d'Eurostat*. [en ligne]. Juin 2025. Statistics Explained. [Consulté le 24 octobre 2025]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/statistics-explained/index.php?oldid=545642*.

- FELLEGI, Ivan P. and HOLT, Daniel T., 1976. A Systematic Approach to Automatic Edit and Imputation. In : *Journal of the American Statistical Association*. Mars 1976. Taylor & Francis Ltd.. Volume 71, n° 353, pp. 17-35.

- FERMOR-DUNMAN, Verena et PARSONS, Laura, 2022. Data Acquisition Processes Improving Quality of Microdata at the Office for National Statistics. In : *European Conference on Quality in Official Statistics (Q2022)*. [en ligne]. 8 juin – 10 juin 2022. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://q2022.stat.gov.lt/scientific-information/papers-presentations/session-25*.

- FORTEZA, Nicolás et GARCÍA-URIBE, Sandra, 2025. A Score Function to Prioritize Editing in Household Survey Data: A Machine Learning Approach. In : *Journal of Official Statistics*. [en ligne]. Mars 2025. Volume 41, n° 1, pp. 144-171. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://journals.sagepub.com/doi/10.1177/0282423X241309971*.

- GITELMAN, Lisa, 2013. *"Raw Data" Is an Oxymoron*. The MIT Press. ISBN 978-0-262-51828-4.

- GRANQUIST, Leopold, 1997. The New View on Editing. In : *International Statistical Review*. [en ligne]. Décembre 1997. International Statistical Institute. Volume 65, nº 3, pp. 381-387. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.1997.tb00315.x*.

- GRANQUIST, Leopold et KOVAR, John G., 1997. Editing of Survey Data: How Much Is Enough?. In : *Survey Measurement and Process Quality*. Février 1997. John Wiley & Sons, Inc.. Pp. 415-435. ISBN 978-0-471-16559-0.

- HAND, David J., 2018. Statistical Challenges of Administrative and Transaction Data. In : *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. [en ligne]. Juin 2018. Volume 181, nº 3, pp. 555-605. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://academic.oup.com/jrsssa/article/181/3/555/7072022?login=false*.

- 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. Nº N1, pp. 25-34. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/3647025?sommaire=3647035*.

- INSEE, 1998. Mesurer la pauvreté aujourd'hui. In : *Économie et statistique*. [en ligne]. Octobre 1998. Nº 308-310. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.persee.fr/issue/estat_0336-1454_1998_num_308_1?sectionId=estat_0336-1454_1998_num_308_1_2588*.

- JOUBERT, Marie-Pierre, 2025. Les données de téléphonie mobile – Une source de connaissance sur la population et ses déplacements. In : *Courrier des statistiques*. [en ligne]. 23 juin 2025. Insee. Nº N13, pp. 95-118. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8546945?sommaire=8546949*.

- KOUMARIANOS, Heidi, RIVIÈRE, Pascal, 2025. Statistiques fondées sur des données administratives : esquisse d'un cadre général. In : *Documents de travail.* [en ligne]. 23 juin 2025. Insee. Nº M2025/03. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/8596376*.

- LAMARCHE, Pierre et LOLLIVIER, Stéfan, 2021. Fidéli, l'intégration des sources fiscales dans les données sociales. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2021. Insee. Nº N6, pp. 28-46. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5398683?sommaire=5398695*.

- LAWRENCE, David, et McKENZIE, Richard, 2000. The General Application of Significance Editing. In : *Journal of Official Statistics*. [en ligne]. Janvier 2000. Volume 16, nº 3, pp. 243-253. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.researchgate.net/publication/256199660_The_General_Application_of_Significance_Editing*.

- 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. Nº N3, pp. 61-75. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254225?sommaire=4254170*.

- LEFEBVRE, Olivier, 2024. Le Répertoire Statistique des Individus et des Logements (Résil) – Un nouvel univers de référence pour les statistiques démographiques et sociales. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2024. Insee. N° N11, pp. 73-94. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8203040?sommaire=8203072*.

- LESUR, Romain, 2025. Sources de données privées : panorama et perspectives. In : *Courrier des statistiques*. [en ligne]. 23 juin 2025. Insee. N° N13, pp. 73-94. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8546943?sommaire=8546949*.

- PANNEKOEK, Jeroen, SCHOLTUS, Sander, et VAN DER LOO, Mark, 2013. Automated and Manual Data Editing: a View on Process Design and Methodology. In : *Journal of Official Statistics*. [en ligne]. Novembre 2013. Volume 29, n° 4, pp. 511-537. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://journals.sagepub.com/doi/abs/10.2478/jos-2013-0038*.

- PARKINSON, Cyril Northcote, 1955. Parkinson's Law. In : *The Economist*. 19 novembre 1955. The Riverside Press. Volume 19.

- RENNE, Catherine, 2018. Bien comprendre la déclaration sociale nominative pour mieux mesurer. In : *Courrier des statistiques*. [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 35-44. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/3647029?sommaire=3647035*.

- RUBIN, Donald B., 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons Inc.. ISBN 0-471-65574-0.

- STIGLITZ, Joseph E., SEN, Amartya et FITOUSSI, Jean-Paul, 2009. *Vers de nouveaux systèmes de mesure – Performances économiques et progrès social*. Odile Jacob. ISBN 978-2-7381-2463-0.

- TAVERNIER, Jean-Luc, 2020. Fonctionnement de l'Insee dans la période de confinement. In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. Insee. N° N5, pp. 6-20. [Consulté le 15 septembre 2025]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5008681?sommaire=5008710*.