N°5

COURRIER DES STATISTIQUES

Décembre 2020

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE

Rédaction en chef
Odile Rascol
Contribution
Insee : Marc Christine, Valérie Darriau, Pascal Rivière, Nicole Roth, Jean-Luc Tavernier
Autorité de la statistique publique : Dominique Bureau
Cnav : Bryan Bellanger, Samuel Goujon
DEPP : Franck Evain
Directeur de la publication
Jean-Luc Tavernier
Directeur de la collection
Pascal Rivière

Rédaction
Maryse Cadalanu, Pierre Glénat, Fabienne Le Hellaye, Odile Rascol, Pascal Rivière
Composition
Agence LATITUDE Nantes
5, rue Jacques Brel
« Les Reflets » Bâtiment A
44 800 SAINT-HERBLAIN
02 51 25 06 06
www.agence-latitude.fr
0425/20
Photo de couverture
Adobe Stock®

Éditeur
Institut national de la statistique et des études économiques
88, avenue Verdier
92 541 MONTROUGE CEDEX

www.insee.fr

Courrier des statistiques N5

SOMMAIRE

PRÉSENTATION DU NUMÉRO

Même si le *Courrier des statistiques* n'a pas vocation à être adhérent à l'actualité, ce deuxième numéro de l'année 2020 ne pouvait pas passer à côté : comment fonctionne un institut national statistique en temps de crise sanitaire ? Au-delà du fonctionnement interne à l'Insee, forcément particulier en période de confinement, il était important que son directeur général **Jean-Luc Tavernier** témoigne de la manière dont l'institut s'est adapté pour jouer pleinement son rôle dans le débat public : produire de nouvelles statistiques, innover dans l'exploitation de nouvelles sources, dans les méthodes d'élaboration du diagnostic conjoncturel, faire preuve d'agilité. Ces maîtres mots des orientations stratégiques de l'Insee ont pris une saveur particulière en 2020.

Avec les deux articles suivants, le *Courrier* s'intéresse à des sujets structurants pour la gouvernance du service statistique public (SSP) français. **Dominique Bureau** revient ainsi sur l'Autorité de la statistique publique, qu'il préside, et dresse un bilan des dix années écoulées depuis sa création. Avec l'ASP, la tradition d'indépendance du SSP, tant dans la production que la diffusion de ses données, bénéficie d'un cadre réglementaire conforme aux engagements européens, et d'une instance chargée de contrôler son application. Cette exigence s'étend désormais au champ des statistiques publiques élaborées par d'autres organismes ayant des missions de service public. **Nicole Roth** et **Marc Christine** détaillent le rôle du Comité du label dans la labellisation de ces statistiques. Ils rappellent aussi le mode d'intervention du Comité sur le champ plus traditionnel des enquêtes de la statistique publique. L'ASP et le Comité du label travaillent ainsi de concert, pour garantir le respect des standards de qualité méthodologique et de transparence du Code de bonnes pratiques européen. À l'ère du numérique et de l'explosion des *data*, les défis de la « qualification » des statistiques prennent toute leur importance.

L'article suivant s'intéresse autant aux données qu'à une manière de les valoriser : le carroyage. Cette technique permet une visualisation cartographique originale de statistiques issues de données individuelles. **Valérie Darriau** décrit comment le chargé d'études à l'Insee, comme le cartographe dans une collectivité locale, peuvent ainsi se placer au plus près des besoins d'études des acteurs du débat public local. L'expérience française ouvre des perspectives mais soulève aussi de redoutables problématiques d'interprétation et de garantie du secret.

L'article de **Franck Evain** sur les indicateurs de valeur ajoutée des lycées, se penche sur la manière dont un service statistique ministériel travaille sa communication pour améliorer le « bon usage » des indicateurs qu'il produit. Car quand la réponse aux besoins de pilotage interne d'une administration débouche sur le développement de données d'évaluation, ces données intéressent alors aussi bien le citoyen que les médias : l'exemple des IVAL permet de mesurer les écueils d'une communication mal maîtrisée, et les avantages à prendre les devants dans ce domaine.

Le numéro N4 avait largement développé la thématique des modèles de microsimulation. L'article de **Bryan Bellanger** et **Samuel Goujon** apporte une nouvelle contribution à la connaissance de ces modèles de plus en plus utilisés. Avec Prisme, le modèle dynamique développé à la Cnav sur les régimes de retraite, les auteurs ne se limitent pas à une présentation technique : ils abordent les atouts d'un modèle développé au sein d'un régime de retraite avec la proximité de la sphère métier. Le modèle, qui a été progressivement

étendu à tous les régimes de retraite, permet aujourd'hui d'éclairer le législateur en période de réforme, ce qui était encore récemment à l'ordre du jour en France, jusqu'à ce qu'une autre actualité nous rattrape.

Enfin, **Pascal Rivière** revient sur une question au cœur du métier de statisticien, en apparence simple : qu'est-ce qu'une donnée ? Pour travailler efficacement des données, il faut savoir comment les caractériser, les appréhender dans leur environnement, et plus généralement ne pas omettre d'explorer chaque facette de ce matériau si divers. Avec cet article, l'auteur nous encourage à prendre de la distance avec l'objet du quotidien des statisticiens.

Odile Rascol
Rédactrice en chef, Insee

FONCTIONNEMENT DE L'INSEE DANS LA PÉRIODE DE CONFINEMENT

*Jean-Luc Tavernier**

***Note de l'éditeur** Il semblait difficile de produire un nouveau numéro du* Courrier des statistiques *sans évoquer l'impact de la pandémie sur les statistiques publiques : cette période a été riche en innovations, particulièrement lors du premier confinement, au printemps 2020. Le présent article trouve son origine dans un premier papier écrit par le directeur général de l'Insee pour le* Journal of the IAOS, *faisant le point sur les nouvelles sources, méthodes et organisations mises en place à cette occasion. Il s'est enrichi de nombreuses informations, notamment sur les publications qui en ont découlé :* nowcasting *publié bimensuellement de fin mars à mi-juillet, statistiques de décès, données de synthèse sur les mobilités. Par souci de clarté, le choix a été fait de se limiter aux développements de cette période et aux analyses qui en résultèrent, jusqu'à septembre 2020. La période qui a suivi a soulevé d'autres difficultés, non évoquées ici, par exemple les modalités de poursuite de la coopération avec les différents opérateurs et la question de la gratuité.*

L'article est écrit à la première personne : il a paru important de présenter la démarche adoptée, les décisions prises, mais aussi de façon plus opérationnelle, les impacts sur l'organisation de l'activité statistique au quotidien.

***Editor's note** It seemed difficult to produce a new issue of the* Courrier des statistiques *without mentioning the impact of the pandemic on official statistics, especially since this period was rich in innovations, particularly during the first lockdown (Spring 2020). This article is based on a first paper written by the Director General of INSEE for the* Journal of the IAOS, *presenting the new sources, methods and organisational elements put in place on this occasion. It has been enriched with a great deal of information, in particular on the resulting publications: nowcasting published every two weeks from late March to mid-July, death statistics, summary data on mobility. For the sake of clarity, we chose a scope : original developments of the spring of 2020, and resulting analyses up to September 2020. The period that followed raised other difficulties, such as the modalities of continuing cooperation with the various operators , and the issue free of charge* vs *paying, but these will not be discussed here.*

Finally, unlike the other articles in the Courrier, this one is written in the first person: it seemed important to present the approach adopted, the decisions taken, but also, in a more operational way, the impacts on the organisation of day-to-day statistical activity.

* Directeur général de l'Institut national de la statistique et des études économiques (Insee), *jean-luc.tavernier@insee.fr*

Dès le début de la période de confinement[1] en France, l'Insee a réussi à réorganiser son travail pour assurer la continuité de ses missions : cela grâce à la généralisation du télétravail, et à l'adaptation de certaines enquêtes auprès des ménages, dont la collecte est passée du face-à-face aux entretiens téléphoniques.

L'Insee a également fourni de nouveaux résultats utiles aux décideurs et au public. En utilisant de nouvelles sources de données – telles que les transactions par carte de crédit ou les données des téléphones mobiles –, et de nouvelles méthodes – principalement le *nowcasting –,* l'institut a permis d'éclairer trois sujets d'intérêt : l'évolution de la situation économique française en temps réel, la répartition de la population présente sur l'ensemble du territoire national et l'évolution du taux de mortalité.

LA CONTINUITÉ DU SERVICE PAR TEMPS DE CONFINEMENT

Dès l'apparition des premiers foyers épidémiologiques[2] en France début mars 2020, l'Insee a pris la décision de suspendre la collecte sur le terrain dans ces zones, qu'il s'agisse des enquêtes en face-à-face auprès des ménages ou des relevés de prix dans les magasins. Dans le reste du pays, en accord avec les organisations syndicales de l'Insee, la collecte était maintenue, avec la consigne de ne pas insister si l'enquêté était réticent ou s'il présentait des symptômes de la maladie. Cependant, ces dispositions n'ont vécu que quelques jours car tout s'est accéléré brutalement à la mi-mars avec la perspective d'un confinement de plusieurs semaines.

Fort d'une expérience de quelques années dans le développement du télétravail, l'Insee a pu compléter rapidement l'équipement des agents en postes nomades : le lundi 16 mars, une majorité des agents en étaient équipés (de l'ordre de 4 000 portables pour 6 000 personnes, y compris enquêteurs).

Lors du dernier comité de direction qui s'est tenu au siège de la Direction générale, ce même 16 mars, les principes généraux suivants sont établis :

- jusqu'à la suspension de l'urgence sanitaire, le télétravail devient la règle absolue, et seuls des cas de force majeure justifient encore une présence physique dans les bâtiments. De fait, les locaux de l'Insee, à Paris-Montrouge et en région, seront vides à compter du 18 mars, avec seulement quelques visites épisodiques pour vérifier l'état du bâtiment ou récupérer du courrier ;
- pour le bon fonctionnement du travail à distance et la répartition des connexions aux serveurs, il est décidé d'en restreindre l'accès au départ. La direction du Système d'information estime que les serveurs peuvent supporter 2 800 accès simultanés, je décide de les limiter à 2 000, limite qui sera élargie au fil des semaines, grâce à l'installation de nouveaux serveurs et à un pilotage rigoureux des infrastructures et des réseaux. Au début de la période de confinement, la plupart des agents devront respecter des plages de connexion (le matin ou l'après-midi selon les cas). Celles-ci sont interrompues chaque jour, entre 12h30 et 13h30, pour permettre aux informaticiens de pratiquer en toute quiétude les maintenances et les montées en régime nécessaires ;

1. L'article est tiré d'un papier écrit pour le *Statistical Journal of the IAOS* à la sortie de la première période de confinement (17 mars-11 mai 2020)(Tavernier, 2020b). Depuis la France est entrée dans un deuxième confinement, le 30 octobre 2020, et y était encore à la date de publication de ce numéro du Courrier des statistiques.
2. Les « clusters ».

- il est demandé aux managers de proximité de garder un contact avec tous les agents, même ceux qui ne sont pas équipés en poste nomade et qui sont invités à consulter leur messagerie sur leur ordinateur personnel ;
- le comité de direction se réunira en audioconférence (que nous avons jugée aussi efficace et moins gourmande en bande passante que la visioconférence) deux à trois fois par semaine. La secrétaire générale rencontrera les organisations syndicales en audioconférence au moins une fois par semaine. Le directeur en charge de l'animation du réseau des établissements régionaux assurera l'échange d'informations avec les directeurs régionaux, à raison d'une courte audioconférence quotidienne ;
- les enquêtes en face-à-face sont suspendues, et basculées lorsque c'est possible en enquêtes par téléphone.

Au 16 mars, la poursuite des relevés de prix dans certains magasins restait en débat. En effet, depuis janvier 2020, l'Insee produit une partie de l'indice des prix à la consommation avec des données de caisses[3]. Les équipes en charge de la mesure des prix de détail souhaitaient le maintien de collecte en magasin pour compléter les données recueillies par les enseignes de la grande distribution.

COMMUNIQUER EN TEMPS DE CRISE, EN EXTERNE...

« Un communiqué de presse est diffusé le 16 mars, pour présenter la manière dont l'Insee prévoit d'assurer ses missions. »

Un communiqué de presse est diffusé le 16 mars, pour présenter la manière dont l'Insee prévoit d'assurer ses missions (Insee, 2020d). Il affiche une ambition : que l'essentiel des productions et publications de l'Insee soient maintenues. Il signale cependant que certaines activités seront prioritaires : la tenue des registres (des personnes physiques et des entreprises), la production des enquêtes de conjoncture, des indicateurs de court terme et des comptes nationaux, l'exploitation de l'enquête Emploi qui sera entièrement collectée par téléphone. Le communiqué annonce également que la *Note de conjoncture* prévue fin mars est ajournée (cf. *infra*) et que l'Insee se donne pour objectif de publier un état simplifié de la situation économique française deux fois par mois. Enfin, il alerte sur le risque d'une qualité dégradée des statistiques qui sera donc systématiquement commentée.

Ce communiqué a été transmis, une heure avant sa publication, à la Directrice générale d'Eurostat, au directeur du cabinet du ministre français de l'économie, aux présidents de l'Autorité de la statistique publique et du Conseil national de l'information statistique. Il sera envoyé aux chefs des services statistiques ministériels. Une version en anglais est envoyée aux homologues des autres instituts nationaux de statistique.

Une mise à jour le 25 mars annonce la suppression de certaines productions : l'enquête auprès des ménages sur le recours aux technologies de l'information et de la communication (TIC) n'aura pas lieu en 2020, la publication d'indices de prix propres à Mayotte est suspendue.

3. [N.D.L.R.] Voir à ce sujet l'article de Marie Leclair « Utiliser les données de caisses pour le calcul de l'indice des prix à la consommation » paru dans le numéro N3 du Courrier des statistiques (Leclair, 2019).

Le communiqué du 25 mars (Insee, 2020a)[4] signale aussi que les entreprises non répondantes ne seront pas relancées systématiquement ni poursuivies, que les relevés de prix dans les magasins sont provisoirement interrompus. Enfin, il présente des productions statistiques spécifiques faites pendant la crise, et qui feront l'objet de la seconde partie de cet article.

... ET EN INTERNE

Le premier communiqué de presse et les suivants seront également utilisés pour informer les organisations syndicales de l'Insee, et plus généralement, pour communiquer à tous les agents les dispositions prises par l'Insee pour assurer la continuité du service.

En complément, j'adresse un message à tous les agents deux jours après le début du confinement. Je relaie une lettre du ministre de l'Économie et des Finances, auquel l'Insee est rattaché, lettre qui, deux semaines après le début du confinement, nous remercie des dispositions prises et du travail accompli. Et, les agents reçoivent régulièrement des informations de la part de la secrétaire générale et du responsable des Ressources humaines.

UN FONCTIONNEMENT STABILISÉ APRÈS DEUX SEMAINES

Je crois pouvoir dire que tout cela a permis de maintenir un fonctionnement de qualité, quand bien même nous n'avons plus tenu de réunion physique pendant plus de deux mois. Le système d'information ne nous a jamais fait défaut, les collectifs de travail sont restés en contact. Tous les objectifs affichés ont été respectés.

> *« Le système d'information ne nous a jamais fait défaut, les collectifs de travail sont restés en contact. Tous les objectifs affichés ont été respectés. »*

Les premiers comités de direction de la période de confinement ont été dédiés, naturellement, à la gestion de crise. Et, après deux semaines, nous avons réussi à retrouver le temps et la disponibilité pour traiter les sujets usuels, comme les recrutements et nominations.

Il faut aussi souligner que le bureau de presse n'a jamais cessé de fonctionner et qu'il n'y a eu aucune rupture dans la diffusion des communiqués, leur relais sur les réseaux sociaux - j'ai moi-même eu plus de retours que jamais sur LinkedIn - et la réponse aux journalistes et aux *fact checkers*.

Plus d'un mois après le début du confinement, la qualité des informations collectées (tant des enquêtes que des fichiers administratifs) pose davantage de problème que la capacité de l'institut à les traiter. En effet, le taux de réponse de plusieurs enquêtes auprès des entreprises s'est très fortement dégradé, et un certain nombre de déclarations administratives, notamment fiscales, vont sans doute être perturbées. La collecte des prix est loin d'être stabilisée. Côté ménages, le basculement du face-à-face sur le téléphone entraîne des difficultés, notamment pour les enquêtes complexes que sont la première interrogation à l'enquête Emploi et l'enquête sur les conditions de vie ; les taux de réponse

4. [N.D.L.R.] Toutes les mentions de références bibliographiques ont été ajoutées à l'article original.

sont aussi affectés pour les enquêtes ménages, de façon toutefois moins prononcée et moins générale que pour les enquêtes auprès des entreprises.

La pertinence de certaines statistiques risque aussi d'être mise à l'épreuve. Dès la mi-mars, j'ai signalé à mes homologues que les critères du chômage au sens du BIT allaient poser problème. Autre exemple, la pondération utilisée pour calculer les prix à la consommation va également s'écarter de la structure de la consommation pendant la période de confinement.

LES PRODUCTIONS STATISTIQUES EN RÉPONSE À LA CRISE : *NOWCASTING*…

> *« L'Insee présente la particularité de réaliser des prévisions à court terme. »*

Parmi les instituts statistiques, l'Insee présente la particularité de réaliser des prévisions à court terme, c'est-à-dire d'utiliser les enquêtes de conjoncture et des outils de modélisation macroéconomique pour prévoir les agrégats des comptes nationaux pour le trimestre en cours et le trimestre suivant. C'est l'objet des *Notes de conjoncture,* publiées quatre fois l'an.

La *Note* annoncée pour le 24 mars 2020 devait porter sur les prévisions de croissance pour les deux premiers trimestres de 2020. L'exercice, déjà rendu difficile par le développement de l'épidémie en Chine, apparaît impossible avec l'émergence des premiers *clusters* en Europe.

Le 12 mars, la décision est prise de surseoir à la *Note* prévue et de réorienter le travail des conjoncturistes de l'Insee aux fins de pur *nowcasting* : mesurer au mieux, à chaque instant, la chute de l'activité économique.

Le 16 mars, le communiqué de presse signale que les premiers dépouillements des enquêtes de conjoncture de mars montrent que la chute du climat conjoncturel est plus rapide encore qu'à l'automne 2008. Il indique que la *Note de conjoncture* du 24 mars est ajournée, et que l'Insee se donne pour objectif de présenter son appréciation de l'économie toutes les deux semaines. Plus précisément, je demande aux équipes d'estimer la chute du PIB et la chute de la consommation.

La première estimation de ce type paraît le 26 mars, en même temps que les résultats des enquêtes de conjoncture auprès des entreprises. Comme la plupart des réponses des entreprises ont été reçues au début du mois de mars, les données collectées ne reflètent pas encore l'effet des décisions de fermeture des écoles, puis de fermeture des hôtels, cafés, restaurants, enfin de confinement généralisé. De ce fait, la méthode habituelle d'étalonnage[5] sur les enquêtes aurait minoré de beaucoup la chute de l'activité. Il a donc fallu recourir à des méthodes très différentes.

La méthode la plus naturelle a consisté à recueillir de l'information de la part d'entreprises ou de branches professionnelles, directement ou par le truchement de partenaires, par exemple la Banque de France ou un institut de conjoncture proche des milieux patronaux.

5. [N.D.L.R.] Les étalonnages à l'aide d'enquêtes de conjoncture constituent un outil essentiel de prévision à court terme pour la production manufacturière et le PIB (Dubois et Michaux, 2006).

Cette information a été traitée, secteur par secteur, à un niveau de désagrégation très fin (138 postes). Une première compilation donnait un ordre de grandeur de perte d'activité à hauteur d'un tiers du PIB.

Mais j'ai souhaité qu'on ne se repose pas sur cette seule méthode et qu'on corrobore ce résultat avec des sources de données effectives, disponibles à haute fréquence. En la matière, les conjoncturistes pensent en premier lieu à la consommation d'électricité, mais d'autres pistes ont été testées et rapidement écartées (indicateurs de pollution, *Google Trends*, vocabulaire utilisé dans la presse, etc.). Nous avons privilégié les statistiques issues des transactions de cartes bancaires, auxquelles le Groupement des Cartes Bancaires CB qui rassemble les principaux réseaux bancaires nous a donné exceptionnellement accès.

… FACILITÉ PAR LES CONTACTS NOUÉS AUPARAVANT…

À vrai dire, nous avions engagé des discussions avec le Groupement des Cartes Bancaires CB depuis quelque temps, et nous étions sur le point de lancer deux ou trois projets pour tester l'intérêt des données de cartes de crédit. Cela était venu dans l'actualité de notre réflexion statistique des derniers mois dans le contexte d'une grande grève prolongée des transports en commun en décembre 2019, pour mesurer de façon rétrospective la chute de l'activité des commerces parisiens.

À la mi-mars, l'intervention de la fédération professionnelle des banques mais aussi celle, à mon initiative, du Président d'une grande banque française permettront d'accélérer la coopération. Le 18 mars, l'Insee signe ainsi une convention avec le Groupement des Cartes Bancaires CB pour disposer, pendant la durée de la crise, des données quotidiennes de transactions, désagrégées par produit. Nous avons aussi obtenu les données qui nous permettent d'avoir un recul historique de deux années. En quelques jours, l'exploitation de cette nouvelle source a permis de confirmer l'ordre de grandeur d'une chute d'un tiers pour la consommation – après des « hoquets » liés au comportement de stockage des ménages, consécutif aux annonces de confinement. Nous avons pu obtenir par ailleurs des résultats directement utilisables en matière de valeur ajoutée dans certains secteurs de services aux ménages pour lesquels les remontées microéconomiques s'avéraient lacunaires.

> *« Le 26 mars, l'Insee publie donc une estimation de chute de PIB de 35 % et de chute de la consommation des ménages du même ordre de grandeur. »*

Le 26 mars, l'Insee publie donc une estimation de chute de PIB (en instantané par rapport à un régime normal) de 35 % et de chute de la consommation des ménages du même ordre de grandeur (Insee, 2020e). La publication ***(figure 1)*** est précédée d'un avant-propos dans lequel je signale, d'une part qu'il peut paraître incongru de parler économie alors que le paroxysme de la crise sanitaire n'est pas encore atteint, d'autre part que l'ordre de grandeur est nécessairement fragile ; mais j'ajoute que j'estime, grâce notamment à l'exploitation des données de transactions électroniques, que cet ordre de grandeur paraît suffisamment robuste pour qu'il puisse être publié.

Je pense que c'est la première estimation de ce type qui ait été réalisée et publiée par un institut statistique depuis le début de la crise sanitaire, sans doute aussi précurseur de l'exploitation de « *hard data* ».

Il me semble important de signaler que cette publication a suivi le processus usuel, à savoir une diffusion sous embargo au seul directeur du cabinet du ministre de l'Économie et des Finances, quelques heures avant la publication le 26 mars à 7h30 (l'application au cas français des règles de bonnes pratiques de la statistique publique). La publication du 26 mars a été mise à disposition en anglais (Insee, 2020c), et diffusée à mes homologues.

Depuis, l'ordre de grandeur s'est révélé assez fiable : il a été confirmé par deux mises à jour bimensuelles : le 9 avril, jour où je l'ai présenté à la Commission des finances de l'Assemblée nationale, et le 23 avril. La Banque de France et des instituts de conjoncture nationaux ont diffusé, depuis le 26 mars, des évaluations proches[6].

Figure 1. Première estimation de la chute du PIB

...parue dans le point de conjoncture du 26 mars 2020

(Insee, 2020 e)

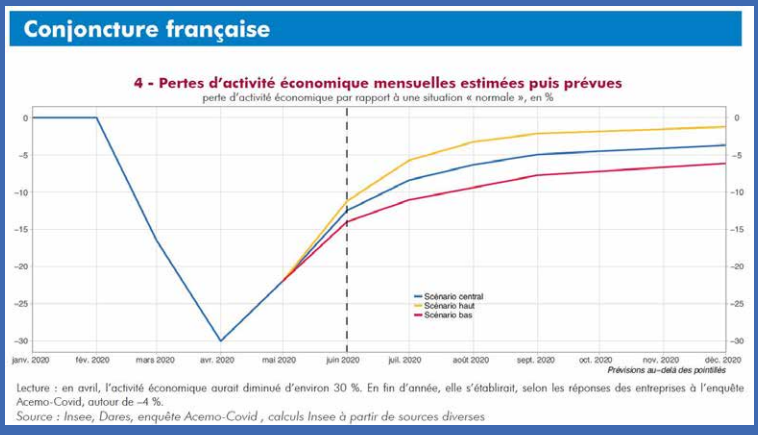
Tableau 1
Estimation de la perte d'activité liée aux mesures d'endiguement (écart entre l'activité économique estimée pendant la dernière semaine de mars et l'activité d'une semaine « normale »)

Branches d'activité	Part dans le PIB (en %)	Hypothèse de perte d'activité par rapport à la normale (en %)	Contribution à la baisse d'activité (en points de PIB)
Agriculture et industries agro-alimentaires	4	- 4	0
Industrie hors agro-alimentaire	12	- 52	- 6
Construction	6	- 89	- 5
Services marchands	56	- 36	-20
Services non marchands	22	-14	-3
Total	100	- 35	- 35

Calculs Insee, à partir de sources diverses

... et dans le point de conjoncture du 8 juillet 2020

(Insee, 2020h).



6. [N.D.L.R.] Voir également les trois articles du blog de l'Insee (Blanchet, 2020 ; Pouget, 2020 ; Tavernier, 2020a) sur les enseignements pour l'institut du traitement opéré en temps de crise sur la mesure de l'activité économique.

… POPULATION PRÉSENTE SUR LES TERRITOIRES…

Depuis plusieurs années, l'Insee cherche à travailler avec des opérateurs de téléphonie mobile, notamment pour mesurer quelle est la population présente, dans un territoire donné, à un moment donné. Comme dans le cas des données de cartes de crédit, les discussions étaient rendues longues du fait de considérations juridiques et financières.

Quelques jours après le confinement, il est apparu qu'il y avait eu un départ significatif des résidents des grandes métropoles, notamment de Paris, pour rejoindre leur résidence secondaire. Il m'a semblé que la répartition des Français sur le territoire national ayant été affectée, une information chiffrée devait être d'un certain intérêt pour les services de santé, d'approvisionnement, de police. En discutant avec un Préfet, j'avais d'ailleurs eu confirmation de cet intérêt.

Je m'en suis ouvert à un cadre dirigeant d'Orange, principal opérateur de téléphonie en France, en lui proposant une coopération entre nos *data scientists*. Ayant évidemment reçu d'autres sollicitations, Orange a décidé de partager gratuitement l'exploitation de ses données qui pouvaient intéresser les pouvoirs publics en lien avec la gestion de la crise. Il s'agit bien évidemment d'une exploitation statistique, qui ne pose aucun problème au regard du règlement européen de protection en matière de données personnelles[7]. La collaboration des *data scientists* de l'Insee et d'Orange s'avérera utile, elle aidera notamment Orange à affiner ses estimations et l'Insee à mieux appréhender sa connaissance des données de téléphonie mobile et leur traduction en concepts utiles à la statistique publique.

> *« Une publication de l'Insee parue le 8 avril permet de visualiser les départements dont la population avait diminué ou au contraire s'était accrue suite aux migrations ayant précédé le début du confinement. »*

Au final, une publication de l'Insee parue le 8 avril permet de visualiser les départements dont la population avait diminué ou au contraire s'était accrue suite aux migrations ayant précédé le début du confinement (Insee, 2020f et 2020g). Ce sont ainsi environ 10 % des Parisiens (hors étudiants) qui ont quitté leur domicile pour se « relocaliser » le plus souvent à la campagne ***(figure 2)***.

Du point de vue des politiques publiques, une marge de progrès est encore possible : la maille départementale est trop large pour bien dimensionner les services publics ; et peut-être sera-t-il possible de produire des chiffres à un niveau géographique plus désagrégé, et de poursuivre des coopérations similaires avec d'autres opérateurs de téléphonie mobile. D'un point de vue médiatique, la visibilité d'une telle publication est très importante, et illustre assez bien le potentiel statistique pour des informations d'intérêt commun que présentent les données de téléphonie mobile.

7. [N.D.L.R.] Voir référence juridique en fin d'article.

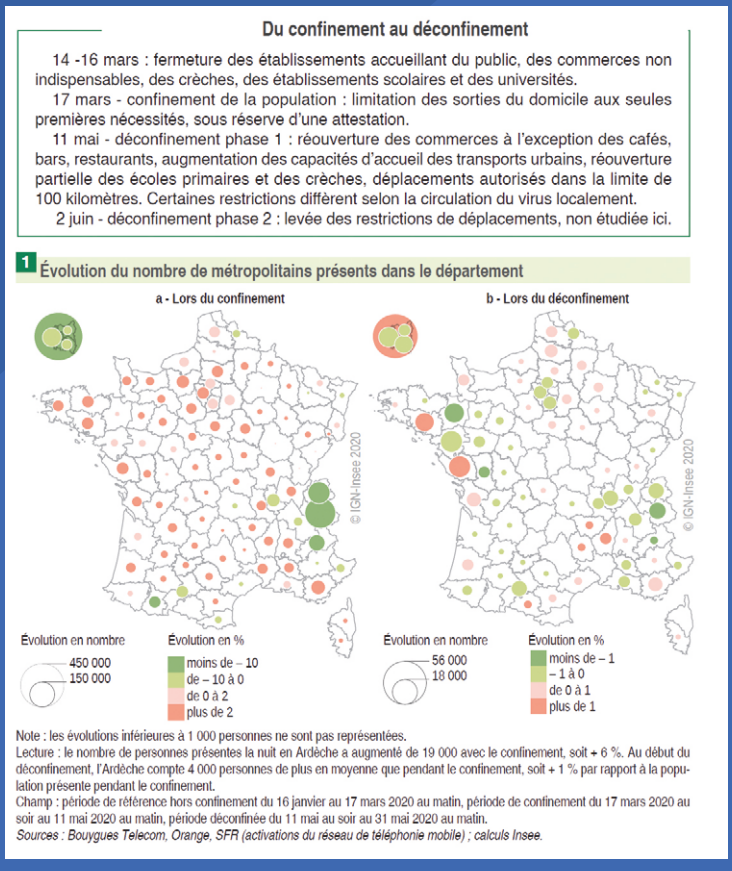
… ET STATISTIQUES DE SURMORTALITÉ

Le troisième effort exceptionnel réalisé par l'Insee est, hélas, moins original puisqu'il a trait à une statistique sur laquelle la plupart des instituts statistiques font actuellement porter leurs efforts : le recensement des décès.

Usuellement, l'Insee publie, vers le 20 du mois, le nombre de naissances et de décès, tels qu'ils sont enregistrés le mois précédent par l'état civil de toutes les communes de France. En temps normal, ces statistiques retiennent assez peu l'attention. Mais il est apparu assez vite, dès la mi-mars, que le décompte des décès à l'hôpital imputés à la Covid-19 ne suffirait pas à rendre compte de la surmortalité.

Depuis le 27 mars, l'Insee publie chaque vendredi les statistiques de décès survenus la semaine antérieure (les communes ont une semaine pour transmettre les données d'état civil à l'Insee). La surmortalité peut être estimée, en comparant le cumul des décès depuis le début du mois de mars au cumul des décès sur les périodes analogues des années

Figure 2. Statistiques sur la mobilité des personnes parues en juillet 2020



(Coudin *et alii*, 2020)

précédentes. Dès le 27 mars, les statistiques étaient données au niveau national, régional et départemental.

La publication s'est enrichie de nouveaux tableaux au fil des semaines. On trouve aujourd'hui des désagrégations, par genre, tranche d'âge et type de lieu de décès[8]. L'intérêt pour ces chiffres est évidemment immense, tant au niveau national que local. Ils sont régulièrement présentés dans les conférences de presse des autorités sanitaires.

> *« Ceux qui mettent à disposition actuellement des statistiques de décès par pays, ne présentent pas systématiquement des chiffres qui sont comparables d'un pays à l'autre. »*

Conformément à la pratique de l'Insee, ces tableaux sont accompagnés de commentaires, qui permettent notamment de bien interpréter la comparaison avec les années précédentes. À titre d'exemple, la grippe hivernale avait été assez virulente en mars 2018, si bien que jusqu'à quasiment la fin mars, au niveau national, on ne constatait pas de surmortalité sensible par rapport à cette année-là. L'approche statistique de la surmortalité a également été présentée dans un article publié sur le tout nouveau blog de l'Insee (Bayet, Le Minez et Roux, 2020a et 2020b).

Je constate, avec une certaine amertume, que ceux qui mettent à disposition actuellement des statistiques de décès par pays, à l'heure où j'écris, ne présentent pas systématiquement des chiffres qui sont comparables d'un pays à l'autre. Il me semble qu'il y a ici un effort important à faire pour l'ensemble de la statistique publique afin de fournir, au niveau international, les statistiques les plus comparables possible, et en tout cas, en explicitant les limites de la comparabilité ***(figure 3)***.

Je n'ai pas évoqué les statistiques de causes de décès. L'information élémentaire, telle qu'elle est enregistrée par les médecins dans les certificats de décès, n'est pas traitée par l'Insee, mais par un organisme de recherche en matière de santé, l'Inserm[9]. Dans ce domaine, en période de crise sanitaire, il faut se poser la question de la fréquence de publication (le règlement européen prévoit une périodicité annuelle[10]). Il faut aussi se poser la question de la comparabilité de l'information collectée selon les pays et les pratiques d'établissement des certificats de décès.

LES ENQUÊTES LANCÉES DE FAÇON EXCEPTIONNELLE

L'Insee a ajouté des questions spécifiques dans l'enquête mensuelle de conjoncture auprès des ménages (enquête Camme par téléphone auprès de 2 000 personnes), afin d'apprécier les effets du confinement sur la vie des ménages ; la première exploitation est sortie en mai[11].

L'Insee a aussi exploité les réponses ouvertes des entreprises dans les enquêtes de conjoncture pour prendre la mesure, par une analyse textuelle, de l'inquiétude générale suscitée par l'épidémie (Insee, 2020i et 2020j).

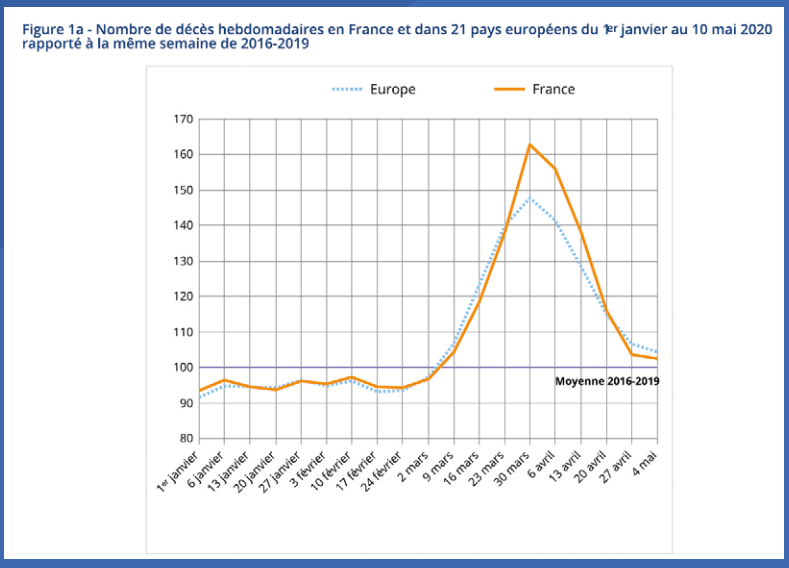
8. [N.D.L.R.] Tels que renseignés par les officiers d'état civil qui enregistrent les actes de décès : établissement hospitalier, maison de retraite, domicile, autre ou non renseigné.
9. [N.D.L.R.] Institut national de la santé et de la recherche médicale.
10. [N.D.L.R.] Avec des délais pouvant aller jusqu'à 18 mois après la fin de l'année de référence.
11. [N.D.L.R.] En fait la première (et seule à ce jour) exploitation est sortie en juin, sur la base du module ajouté à l'enquête Camme de mai (Albouy et Legleye, 2020).

Avec l'aide de l'Insee, la Dares[12], service statistique du ministère du Travail, a adapté l'enquête trimestrielle Acemo (Activité et conditions d'emploi de la main-d'œuvre) menée auprès des entreprises en avril pour avoir des données sur le recours au télétravail, au chômage partiel, etc. L'exploitation de cette enquête a été publiée le vendredi 17 avril (Dares, 2020).

L'Insee prévoit d'adapter une enquête auprès des entreprises prévue en septembre sur la sous-traitance pour évaluer l'impact de la crise sur l'organisation et l'activité des entreprises, sur l'approvisionnement et les chaînes de valeur, etc.[13]

Enfin, l'Insee a été sollicité par différents services et organismes de recherche en matière de santé pour délivrer des échantillons représentatifs. La coopération la plus significative concerne l'Inserm et la Drees[14], service statistique du ministère de la Santé, pour prendre la mesure de la prévalence du virus et de ses symptômes, par une énorme enquête par internet, suivie de tests sur des sous-échantillons. La première vague de cette enquête appelée EpiCov[15], doit démarrer au cours de la dernière semaine d'avril[16].

Figure 3. Statistiques comparatives des décès publiées le 29 juillet 2020



(Dahoo et Gaudy, 2020).

12. [N.D.L.R.] Direction de l'animation de la recherche, des études et des statistiques.
13. [N.D.L.R.] Les résultats de l'enquête ont fait l'objet d'un Insee Première (Duc et Souquet, 2020) en décembre 2020.
14. [N.D.L.R.] Direction de la recherche, des études, de l'évaluation et des statistiques.
15. [N.D.L.R.] Étude ÉPIdémiologique de la diffusion du SARS COV2 (Cnis, 2020 ; Drees, Insee, Inserm et Santé Publique France, 2020a). La première vague a débuté en avril, la deuxième en octobre.
16. [N.D.L.R.] Les résultats de la première vague ont fait l'objet d'une série de publications (Givord, Silhol *et alii*, 2020 ; IRESP, 2020 ; Drees, Insee, Inserm et Santé publique France, 2020b).

EN GUISE DE CONCLUSION PROVISOIRE

Évidemment, toutes ces productions statistiques faites par l'Insee et l'ensemble du système statistique public français, sont très visibles sur le site *web*, elles sont du reste largement commentées. Depuis le 22 avril, elles sont rassemblées, pour plus de visibilité encore, dans une page du site de l'Insee dédiée aux statistiques liées à la Covid-19 (Insee, 2020a).

Pour finir, je ne sais pas si la statistique publique sortira renforcée ou pas de cette épreuve, mais je crois pouvoir dire que l'Insee a fait ses meilleurs efforts, et a tiré parti des circonstances exceptionnelles pour catalyser des coopérations avec les producteurs de données qui étaient en germe depuis longtemps. Je ne saurais trop rendre hommage à l'engagement et à la réactivité des nombreux collègues de l'Insee dans les semaines qui viennent de s'écouler.

Le confinement, le télétravail sont contraignants. Nous craignons tous pour nos proches, nos aînés, nos collègues. C'est sans doute un réconfort malgré tout, dans cette épreuve, que d'exercer une mission qui peut témoigner en temps réel de son utilité. Et pour l'Insee de montrer qu'il est capable de réaliser des productions qu'il n'aurait pas faites sans la crise. Le fait de parler là, dans cet article, de la vie de l'Institut au cours des derniers mois fait sans doute partie de ces choses qu'on n'aurait jamais imaginé faire en temps normal.

BIBLIOGRAPHIE

ALBOUY, Valérie et LEGLEYE, Stéphane, 2020. *Conditions de vie pendant le confinement : des écarts selon le niveau de vie et la catégorie socioprofessionnelle.* [en ligne]. 19 juin 2020. Insee Focus, n°97. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4513259*.

BAYET, Alain, LE MINEZ, Sylvie et ROUX, Valérie, 2020a. Mourir de la grippe ou du coronavirus : faire parler les chiffres de décès publiés par l'Insee… avec discernement In : *Le blog de l'Insee.* [en ligne]. 7 avril 2020. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://blog.insee.fr/mourir-de-la-grippe-ou-du-coronavirus-faire-parler-les-chiffres-de-deces-publies-par-linsee-avec-discernement/*.

BAYET, Alain, LE MINEZ, Sylvie et ROUX, Valérie, 2020b. Statistiques sur les décès : le mode d'emploi des données de l'Insee en 7 questions/réponses. In : *Le blog de l'Insee.* [en ligne]. 14 mai 2020. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://blog.insee.fr/statistiques-sur-les-deces-le-mode-demploi-des-donnees-de-linsee-en-7-questions-reponses/*.

BLANCHET, Didier, 2020. X % de quoi ? Quelle mesure de l'activité pendant la crise, quelle(s) mesure(s) pour l'après-crise. In : *le blog de l'Insee.* [en ligne]. 17 juillet 2020. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://blog.insee.fr/x-de-quoi-quelle-mesure-de-lactivite-pendant-la-crise-quelles-mesures-pour-lapres-crise/*.

CNIS, 2020. EpiCov : Étude EPIdémiologique de la diffusion du SARS-CoV2 – 2020X711SA. In : *site du Cnis*. [en ligne]. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.cnis.fr/enquetes/epicov-etude-epidemiologique-de-la-diffusion-du-sars-cov2-2020x711sa/*.

COUDIN, Élise, DE BELLEFON, Marie-Pierre, GALIANA, Lino, SUARES CASTILLO, Milena et SÉMÉCURBE, François, 2020. *Retour partiel des mouvements de population avec le déconfinement.* [en ligne]. 22 juillet 2020. Insee Analyses, n°54. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4635407*.

DAHOO, Umar et GAUDY, Lisa, 2020. *En France, comme en Europe, un pic de surmortalité lié à la Covid-19 fin mars-début avril.* [en ligne]. 29 juillet 2020. Insee Focus, n°200. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4637552*.

DARES, 2020. *Activité et conditions d'emploi de la main-d'œuvre pendant la crise sanitaire Covid-19.* [en ligne]. 17 avril 2020. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://dares.travail-emploi.gouv.fr/IMG/pdf/dares_acemo_covid19_synthese_17-04-2020.pdf*.

DREES, INSEE, INSERM et SANTÉ PUBLIQUE FRANCE, 2020a. *Site de l'enquête EpiCov.* [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.epicov.fr/*.

DREES, INSEE, INSERM et SANTÉ PUBLIQUE FRANCE, 2020b. *En mai 2020, 4,5 % de la population vivant en France métropolitaine a développé des anticorps contre le SARS-CoV-2.* [en ligne]. 9 octobre 2020. Drees, Études et Résultats, n°1167. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://drees.solidarites-sante.gouv.fr/etudes-et-statistiques/publications/etudes-et-resultats/article/en-mai-2020-4-5-de-la-population-vivant-en-france-metropolitaine-a-developpe*.

DUBOIS, Éric et MICHAUX, Emmanuel, 2006. Étalonnages à l'aide d'enquêtes de conjoncture : de nouveaux résultats. In : *Économie & prévision.* [en ligne]. N°172, 2006-1, pp. 11-28. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.persee.fr/doc/ecop_0249-4744_2006_num_172_1_7477*.

DUC, Cindy et SOUQUET, Catherine, 2020. *L'impact de la crise sanitaire sur l'organisation et l'activité des sociétés.* [en ligne]. 10 décembre 2020. Insee Première, n°1830. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4994488*.

GIVORD, Pauline, SILHOL, Julien *et alii,* 2020. *Confinement : des conséquences économiques inégales selon les ménages.* [en ligne]. 14 octobre 2020. Insee Première, n°1822. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4801313*.

INSEE, 2020a. Les conséquences de la crise sanitaire de la Covid-19 – Impacts économiques, démographiques et sociétaux. In : *site de l'Insee.* [en ligne]. Actualisé le 3 décembre 2020. Communiqués de presse.[Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4479280*.

INSEE, 2020b. Points de conjoncture 2020. In : *site de l'Insee.* [en ligne]. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4473296*.

INSEE, 2020c. Economic outlook 2020. In : *site de l'Insee.* [en ligne]. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/en/statistiques/4473307*.

INSEE, 2020d. Missions de l'Insee dans les semaines à venir. In : *site de l'Insee.* [en ligne]. 16 mars 2020. Communiqué de presse. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4469614*.

INSEE, 2020e. *Point de conjoncture du 26 mars 2020.* [en ligne]. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4473294?sommaire=4473296*.

INSEE, 2020f. Population présente sur le territoire avant et après le début du confinement – Premiers résultats. In : *site de l'Insee.* [en ligne]. 8 avril 2020. Communiqué de presse. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4477356*.

INSEE, 2020g. Population présente sur le territoire avant et après le début du confinement : résultats consolidés. In : *site de l'Insee.* [en ligne]. 18 mai 2020. Communiqué de presse. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4493611.*

INSEE, 2020h. *Point de conjoncture du 8 juillet 2020*. [en ligne]. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4630804?sommaire=4473296*.

INSEE, 2020i. *Commentaire des enquêtes de conjoncture du 23 juillet 2020.* [en ligne]. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4634969/Enquetes-conj230720.pdf*.

INSEE, 2020j. *Commentaire des enquêtes de conjoncture du 27 août 2020.* [en ligne]. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4648170/Enquetes_conj270820.pdf*.

IRESP, 2020. *Les inégalités sociales au temps du COVID-19.* [en ligne]. 9 octobre 2020. Questions de santé publique, n° 40 numéro spécial. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.iresp.net/actualite/numero-speciale-de-la-revue-question-de-sante-publique-les-inegalites-sociales-au-temps-du-covid-19/*.

LECLAIR, Marie, 2019. Utiliser les données de caisse pour le calcul de l'indice des prix à la consommation. In : *Courrier des statistiques.* [en ligne]. 19 décembre 2019. N°N3 , pp. 61-75. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254225/courstat-3-6.pdf*.

POUGET, Julien, 2020. Nouvelles données pour suivre la conjoncture économique pendant la crise sanitaire : quelles avancées ? quelles suites ? In : *Le blog de l'Insee*. [en ligne]. 28 juillet 2020. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://blog.insee.fr/nouvelles-donnees-pour-suivre-la-conjoncture-economique-pendant-la-crise-sanitaire-quelles-avancees-quelles-suites/*.

TAVERNIER, Jean-Luc, 2020a. La statistique publique à l'épreuve de la crise sanitaire. In : *Le blog de l'Insee*. [en ligne]. 6 mai 2020. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://blog.insee.fr/la-statistique-publique-a-lepreuve-de-la-crise-sanitaire/*.

TAVERNIER, Jean-Luc, 2020b. INSEE operations during the lockdown period. In : *Statistical Journal of the IAOS.* [en ligne]. 9 juin 2020. Volume 36, n° 2, pp. 279-284. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji200669*.

RÉFÉRENCES JURIDIQUES

Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données). In : *Journal officiel de l'Union européenne.* [en ligne]. [Consulté le 15 décembre 2020]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A32016R0679*.

L'AUTORITÉ DE LA STATISTIQUE PUBLIQUE

DIX ANS D'ACTIVITÉ, POUR UNE STATISTIQUE INDÉPENDANTE ET DE QUALITÉ

*Dominique Bureau**

L'Autorité de la statistique publique (ASP) a pour mission première de veiller à l'indépendance professionnelle dans la conception, la production et la diffusion des statistiques publiques. Celle-ci est essentielle vis-à-vis des utilisateurs. Sinon, les statistiques produites deviennent douteuses à leurs yeux et les efforts pour en améliorer la pertinence ou la précision deviennent vains. Dix ans après sa création, cet article réévalue les choix effectués alors pour assurer ce contrôle par l'ASP et examine plus généralement son apport à l'évolution du service statistique public. Il met en exergue une gouvernance équilibrée.

L'ASP veille aussi au respect des principes d'objectivité, d'impartialité, de pertinence et de qualité des données produites. Cette extension de sa mission au-delà de ce qui concerne strictement l'indépendance professionnelle, qui avait été voulue par le législateur, s'est révélée très fructueuse, dans un contexte de bouleversement des données auxquelles a accès le public et de ses attentes, qui impose aux responsables statistiques des exigences nouvelles. Tous les acteurs de la statistique publique sont concernés pour relever ces défis. L'ASP y concourt en veillant au respect de l'ensemble des principes du Code de bonnes pratiques de la statistique européenne.

The primary mission of the Official Statistics Authority (ASP) is to ensure professional independence in design, production and dissemination of the official statistics. This is essential for users of these statistics. Otherwise the credibility of the statistics produced becomes questionable and efforts to improve its relevance or accuracy become futile. Ten years after its creation, this article re-evaluates the choices made to implement this control by the ASP and more generally examines its contribution to the development of the Official Statistical System. It highlights balanced governance.

The ASP ensures compliance with the principles of objectivity, impartiality, relevance and quality of the data produced. This extension of its mission beyond which strictly concerns professional independence, which had been wanted by the legislator, has proved to be very fruitful, in a context of upheaval in the public access to data and public expectations, which imposes new requirements on those responsible for the statistics of the. All those involved in official statistics are concerned to meet these challenges. The ASP contributes to this by ensuring respect for all the principles of the European Statistics Code of Practice.

* Président de l'Autorité de la statistique publique (ASP),
Dominique.Bureau@developpement-durable.gouv.fr

La mission de la statistique publique est de mettre à disposition de tous des données de qualité, pour aider à la prise de décision, alimenter les travaux de recherche et éclairer les débats. Pour leur crédibilité, l'indépendance du service statistique public, qui regroupe l'Insee et les services statistiques des ministères, est essentielle. À cette fin, la loi de Modernisation de l'économie du 4 août 2008 a créé l'Autorité de la statistique publique (ASP), avec comme mission de veiller *« à l'indépendance professionnelle dans la conception, la production et la diffusion des statistiques publiques »*, ainsi qu'au *« respect des principes d'objectivité, d'impartialité, de pertinence et de qualité des données produites »*.

Au moment où l'Autorité se mettait en place, Paul Champsaur, qui en avait la responsabilité, avait analysé *a priori* ses enjeux et moyens (Champsaur, 2009). Dix ans après, il est intéressant de réévaluer les choix effectués alors et d'examiner la contribution de l'Autorité à l'évolution du service statistique public. Après avoir rappelé la genèse de celle-ci et analysé son organisation, les différentes facettes de son activité sont passées en revue dans cette perspective.

LA TRADITION D'INDÉPENDANCE DE L'INSEE INSCRITE DANS LE DROIT

En 2004, alors que l'Union européenne avait besoin de statistiques budgétaires fiables, certaines notifications budgétaires nationales avaient été substantiellement révisées en lien avec des alternances politiques. Suite au constat établi par le Conseil pour les affaires économiques et financières (ECOFIN du 2 juin 2004) la Commission avait, le 25 mai 2005, promulgué une recommandation fixant les principes auxquels devraient se conformer les autorités nationales statistiques. Le premier était celui de l'indépendance vis-à-vis des interventions politiques et autres, qui se devait d'être inscrit dans le droit.

À cet égard, les statisticiens européens venus auditer la situation de l'Insee en janvier 2007, déclaraient : *« nous croyons que l'Insee établit et diffuse les statistiques de façon indépendante sans intervention politique, bien que, contrairement à la situation générale des autres instituts nationaux de statistique du système statistique européen, cette indépendance ne soit pas inscrite dans le droit* [...]. *On ne sera pas surpris que nous recommandions vivement que l'indépendance dans le droit soit accordée à l'Insee dès que possible ».*

En effet, l'Insee est une direction du ministère en charge de l'Économie. De même, les services statistiques ministériels (SSM) contribuent au pilotage des politiques publiques de leur ministère et cette activité peut rendre difficile un travail impartial et indépendant. Ainsi, ils peuvent être soumis à des pressions qui les amèneraient, par exemple, à ne pas publier en temps et en heure les statistiques publiques dont ils ont la charge, les responsables ministériels auxquels ils sont rattachés donnant priorité à d'autres travaux ou souhaitant le report de publications qui ne leur paraîtraient pas politiquement opportunes.

À cette même époque, l'Insee fut confronté à une controverse sur les chiffres du chômage au sens du Bureau international du travail (BIT), du fait des divergences d'évolution qui apparaissaient entre ceux issus de la gestion des demandeurs d'emploi en fin de mois (DEFM) par Pôle emploi et ceux de l'enquête Emploi de l'Insee. Alors qu'être inscrit à Pôle emploi ne signifie pas nécessairement être chômeur au regard de critères stricts d'absence d'activité sur une certaine période, de recherche d'emploi et d'acceptabilité d'une offre éventuelle, le grand public et la presse avaient tendance à privilégier les premiers, ce qui

amena un temps à arrêter la publication des chiffres du chômage au sens du BIT issus de l'enquête Emploi.

Ainsi, l'absence d'inscription dans le droit de l'indépendance professionnelle de la statistique publique était alors de nature à entretenir les suspicions, même si celle-ci était établie dans ses pratiques. La loi de Modernisation de l'économie du 4 août 2008 y remédia en créant l'ASP, tout en reconnaissant que l'indépendance professionnelle ne préjugeait pas du statut de l'organisation qui produit les statistiques : les parlementaires ont alors considéré, comme le gouvernement, que le statut de l'Insee et des services statistiques ministériels, parties intégrantes de l'administration française, était compatible avec leur indépendance professionnelle, sous réserve qu'une institution la contrôle rigoureusement.

« Ainsi, l'absence d'inscription dans le droit de l'indépendance professionnelle de la statistique publique était alors de nature à entretenir les suspicions, même si celle-ci était établie dans ses pratiques. »

L'INDÉPENDANCE SE RÉFÈRE AUX MODALITÉS DE PRODUCTION ET DE DIFFUSION

L'indépendance professionnelle est essentielle vis-à-vis des utilisateurs de la statistique. Sinon, les statistiques produites deviennent douteuses à leurs yeux et les efforts pour en améliorer la pertinence ou la précision deviennent vains. Elle est donc reconnue depuis toujours comme essentielle. Comme le soulignait Edmond Malinvaud[1], *« elle a une telle importance qu'il ne faut manquer aucune occasion de rendre la chose plus manifeste encore et qu'il faut au contraire s'astreindre à ignorer les considérations de court terme qui pourraient conduire à prendre un peu de liberté avec indépendance ou déontologie »* (Champsaur, 2009). Il ajoutait : *« Indépendance et déontologie ne se décrètent pas ; elles se construisent sur le long terme par les pratiques des autorités ministérielles, de l'encadrement et du personnel »*.

Peut-être moins évident mais aussi important, l'indépendance professionnelle conditionne l'accès aux données. Ainsi, cette garantie a été un facteur important dans la décision des enseignes de la grande distribution de s'engager dans le projet d'amélioration du calcul de l'indice de prix *via* l'utilisation des données de caisse, la sécurisation de la confidentialité des données étant au cœur de l'expérimentation puis de l'industrialisation du processus (Leclair, 2019). La mobilisation de telles sources de données étant appelée à se développer, il faut anticiper que leur production exigera des interactions plus profondes avec des entreprises ou personnes privées, pour qui les garanties apportées sur l'indépendance seront déterminantes.

Le premier rôle de l'ASP est donc de veiller à cette indépendance en assurant un traitement précoce des écarts ou des polémiques du type de celles rappelées ci-dessus, ou encore que les ministres des Finances, par exemple, ne s'approprient pas abusivement de bons chiffres conjoncturels auxquels ils auraient eu accès sous « embargo ». En effet, l'accès égal de tout un chacun à la statistique publique découle du principe d'indépendance.

1. Directeur général de l'Insee de 1974 à 1987.

Sur ce point, la réglementation est on ne peut plus explicite : l'ASP doit contrôler que *« les modalités de diffusion des publications du service statistique public respectent les principes de neutralité et d'équité de traitement des utilisateurs et veille notamment à leur diffusion séparée, distincte de toute communication ministérielle »*. L'indépendance professionnelle ne se réfère donc pas seulement à l'indépendance dans les techniques et les méthodes retenues pour élaborer les statistiques. Elle s'applique aussi à ses conditions de diffusion, la façon dont est reçue une statistique dépendant à la fois de son contenu et des circonstances dans lesquelles celle-ci est rendue publique : date de publication, commentaires d'accompagnement, statut de celui qui rend cette statistique publique…

Cependant, la mission de l'ASP est plus large. Elle couvre aussi les principes d'objectivité, d'impartialité, de pertinence et de qualité des données produites. En effet, la qualité constituant le principal atout de la statistique publique dans un monde où l'information prolifère, le plus souvent au détriment du bon éclairage des débats et des choix publics, le Parlement avait souhaité, en même temps qu'il reconnaissait l'obligation d'inscrire l'indépendance de la statistique dans le droit, que l'organisme créé pour en contrôler l'application soit aussi chargé de veiller au respect de l'ensemble des principes déterminants pour assurer cette qualité.

UN CADRE RÉGLEMENTAIRE POUR ÉMETTRE DES AVIS QUI SONT PUBLIÉS

Le décret[2] n°2009-250 a précisé les modalités de fonctionnement de l'ASP. Tout d'abord, il établit l'indépendance de son collège de neuf membres : nommés par les présidents des assemblées, le gouvernement, les grands corps d'inspection et hautes juridictions de contrôle, ils ne reçoivent pas de mandat de la part de l'autorité qui les a désignés. Le mandat du président, nommé par décret en conseil des ministres, est non renouvelable.

> *« À ce titre, elle examine les incidents qui mettent directement en cause l'indépendance professionnelle, telles que les ruptures d'embargo ou les controverses sur certains chiffres. »*

Le décret donne à l'ASP le pouvoir d'émettre des avis généraux sur la mise en œuvre des principes que doit appliquer la statistique publique, et des observations à l'égard de toute personne qui ne s'y conformerait pas. À ce titre, elle examine les incidents qui mettent directement en cause l'indépendance professionnelle, telles que les ruptures d'embargo ou les controverses sur certains chiffres. Ces incidents sont traités avec le souci de dégager des règles qui seront reconnues par tous et feront en sorte qu'ils ne se reproduisent plus.

Un large éventail de canaux de saisine est prévu, dont l'auto-saisine. Ce dernier canal domine en pratique sachant que, lorsque l'Autorité est informée d'un problème potentiel ou est sollicitée pour qu'elle exerce cette capacité, celle-ci examine s'il convient ou non d'inscrire le point correspondant à son ordre du jour. De fait, c'est toujours le cas dès lors

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

que la demande apparaît sérieuse car, pour que l'Autorité exerce pleinement sa mission et que ceci soit crédible, tout cas mettant en cause l'indépendance, l'objectivité ou la qualité des statistiques publiques doit pouvoir lui être soumis, puis examiné.

« Il en est ensuite toujours rendu compte au public, spécifiquement s'il apparaît utile de faire connaître immédiatement l'avis de l'Autorité, sinon dans le cadre de son rapport annuel. »

Il en est ensuite toujours rendu compte au public, spécifiquement s'il apparaît utile de faire connaître immédiatement l'avis de l'Autorité, sinon dans le cadre de son rapport annuel. En effet, il n'a pas été jugé opportun de doter l'Autorité d'un arsenal répressif risquant *in fine* d'être peu praticable, compte-tenu des écueils d'une confrontation récurrente et stérile avec les responsables administratifs et gouvernementaux, ou de l'impuissance vis-à-vis de ceux-ci, et lui ôtant sa crédibilité. En revanche, l'ASP peut rendre publics ses avis. Ceci est un pouvoir bien réel eu égard à l'importance que le public accorde aux enjeux associés à l'indépendance et l'objectivité de la statistique.

Enfin, le décret fixe par ailleurs à l'ASP un certain nombre de tâches particulières à accomplir, en lien avec sa mission ***(encadré 1)***.

UNE GOUVERNANCE DE LA STATISTIQUE PUBLIQUE ÉQUILIBRÉE

Le cadre d'organisation de la statistique publique ainsi mis en place ***(figure 1)*** est équilibré :

- les producteurs du service statistique public sont coordonnés par le directeur général de l'Insee ;
- le Conseil national de l'information statistique (Cnis) oriente l'évolution des enquêtes pour répondre aux attentes des utilisateurs ;
- l'ASP veille au respect du principe d'indépendance et à la qualité.

Les tâches « exécutives » sont donc bien séparées de celles externes d'orientation et de contrôle, et celles-ci sont distinguées. À cet égard, le schéma de l'ASP avait été préféré à l'alternative qui aurait placé un organisme de surveillance au sein du Cnis. Cela risquait de créer une confusion des genres entre les activités de concertation et celles de contrôle. En effet, les missions du Cnis et de l'ASP sont bien différentes : la première oriente ; la seconde veille à l'intégrité du système et à son indépendance. Mais le souci de cohérence se traduit notamment par l'obligation pour l'ASP d'auditionner chaque année le président du Cnis et le directeur général de l'Insee.

UNE VISION ÉLARGIE DE L'ENSEMBLE DES ENJEUX DE LA STATISTIQUE PUBLIQUE

Par ailleurs, le fait que le Président du comité du Secret statistique[3] soit membre du collège de l'Autorité permet une vision partagée des enjeux liés à la protection et à l'accès aux données, domaines en évolution rapide du fait de l'évolution des données massives et de l'essor des nouvelles méthodes d'évaluation des politiques publiques.

> *« Un autre élément important est que le rôle de l'ASP couvre l'ensemble du service statistique public (SSP) et des statistiques publiques. »*

Un autre élément important est que le rôle de l'ASP couvre l'ensemble du service statistique public (SSP) et des statistiques publiques[4]. L'homogénéisation de la statistique publique constitue d'ailleurs un fil directeur de son activité, en ligne avec les recommandations de la réglementation européenne. Ceci n'a pas empêché un renforcement parallèle du rôle de coordination de l'Insee, au contraire, matérialisé par la « charte des SSM » élaborée par sa direction en charge de la coordination statistique, par exemple (Insee, 2019).

Enfin, l'ASP a la charge de labelliser les données des opérateurs publics qui pourraient intéresser la statistique publique, ce qu'elle fait en s'appuyant sur l'instruction du même Comité du label de la statistique publique[5] que pour les enquêtes, ce qui manifeste l'homogénéité voulue de toutes les statistiques publiques en termes d'exigence de qualité.

UNE ORGANISATION SEMBLABLE À CELLE DE L'UE…

L'organisation française apparaît assez comparable à celle de l'Union européenne, où, à côté d'Eurostat, figurent des instances ayant des fonctions proches de celles de l'ASP et du Cnis[6].

En revanche, la France demeure l'un des rares pays à avoir mis en place la recommandation du règlement européen n°223/2009 d'instituer un organe national chargé de veiller à l'indépendance professionnelle des producteurs de statistiques européennes. De plus, les autres pays qui y ont recours, comme la Grèce ou le Royaume-Uni, les ont souvent créés dans un contexte de polémique exacerbée sur la confiance que l'on pouvait accorder à leur institut statistique.

3. Le comité du Secret statistique se prononce sur toute question relative au secret en matière de statistiques. Il donne son avis sur les demandes de communication de données individuelles collectées (loi n° 51-711 du 7 juin 1951 sur l'Obligation, la coordination et le secret en matière de statistiques).
4. Voir à ce sujet l'article de Michel Isnard « Qu'entend-on par statistique(s) publique(s) ? » paru dans le numéro N1 du Courrier des statistiques (Isnard, 2018).
5. Voir l'article de Marc Christine et Nicole Roth sur le Comité du label de la statistique publique, dans ce même numéro.
6. Respectivement l'ESGAB *(European Statistical Governance Advisory Board)* et l'ESAC *(European Statistical Advisory Committee)*.

Encadré 1. Missions spécifiques de l'ASP

Autorité de la statistique publique

Un dispositif inscrit dans la loi qui garantit l'indépendance, la qualité et la pertinence de la statistique publique

Journal officiel "Lois et Décrets"
JORF n°0227 du 17 septembre 2020
ELI: *https://www.legifrance.gouv.fr/eli/jo/2020/9/17/0227*
DÉCRETS, ARRÊTÉS, CIRCULAIRES
Textes généraux

Avis

Consultée sur tout **projet de décret portant sur les missions :**
- de l'Insee
- des services statistiques interministériels (SSM)

Autorité de la statistique publique

Avis

Avis du 10 décembre 2019 de l'Autorité de la statistique publique pour le comité d'audition pour la nomination du directeur de la Direction de la recherche, des études, de l'évaluation et des statistiques (DREES).

Émet un avis sur les **nominations** :
- du directeur général de l'Insee
- des chefs de services statistiques ministériels

Autorité de la statistique publique

Avis

Délivre un avis sur tout projet d'arrêté fixant la **liste des services statistiques ministériels**

Autorité de la statistique publique

Label

Avis du 14 avril 2020 de l'autorité de la statistique publique sur le renouvellement de la labellisation des séries trimestrielles des effectifs salariés, de masse salariale, et de déclaration d'embauche au niveau national et sur la labellisation des séries d'effectifs salariés et de mesure salariale localisées produites par l'Agence centrale des organismes de sécurité sociale (ACOSS).

Labellise, avec l'appui du Comité du label, les séries qui intéressent la statistique publique et produites par :
- des opérateurs publics
- des organismes privés chargés d'une mission de service public

Autorité de la statistique publique

Rapport annuel

Rédige chaque année un rapport public pour le Parlement sur :
- l'exécution du programme de travail de la statistique publique
- la conformité au code de bonnes pratiques de la statistique européenne

Voir (ASP, 2020).

… BIEN QU'ORIGINALE, COMPARÉE AUX AUTRES PAYS

Certains responsables statistiques nationaux font valoir que, chez eux, la création d'une instance comme l'ASP, au lieu de défendre le système statistique public, constituerait un geste de défiance envers celui-ci.

Pragmatisme ou politique de l'autruche ? Ceci est à apprécier au cas par cas. Cependant, il faut souligner, qu'évidemment, l'objectif n'est pas de créer des suspicions non fondées mais bien de renforcer la confiance dans la statistique. À cet égard, toute cartographie des risques pesant sur les systèmes statistiques souligne à quel point l'indépendance professionnelle porte des risques majeurs en cas de remise en cause. Dans cette perspective, un garant externe comme l'est l'ASP est précieux, ne serait-ce que parce qu'il constitue un rempart contre les malentendus ou les polémiques inutiles.

C'est aussi un moyen d'accélérer le règlement de situations insatisfaisantes (car il en demeure forcément), difficiles à réformer sans pression externe obligeant à agir. La remise à plat des statistiques de la pêche, imposée par l'ASP après avoir auditionné en octobre 2014 le service statistique ministériel de la pêche et de l'aquaculture, alors situé au sein de la direction des pêches maritimes et de l'aquaculture, et considéré que le maintien de ce statut à ce service n'était pas souhaitable, en est un bon exemple.

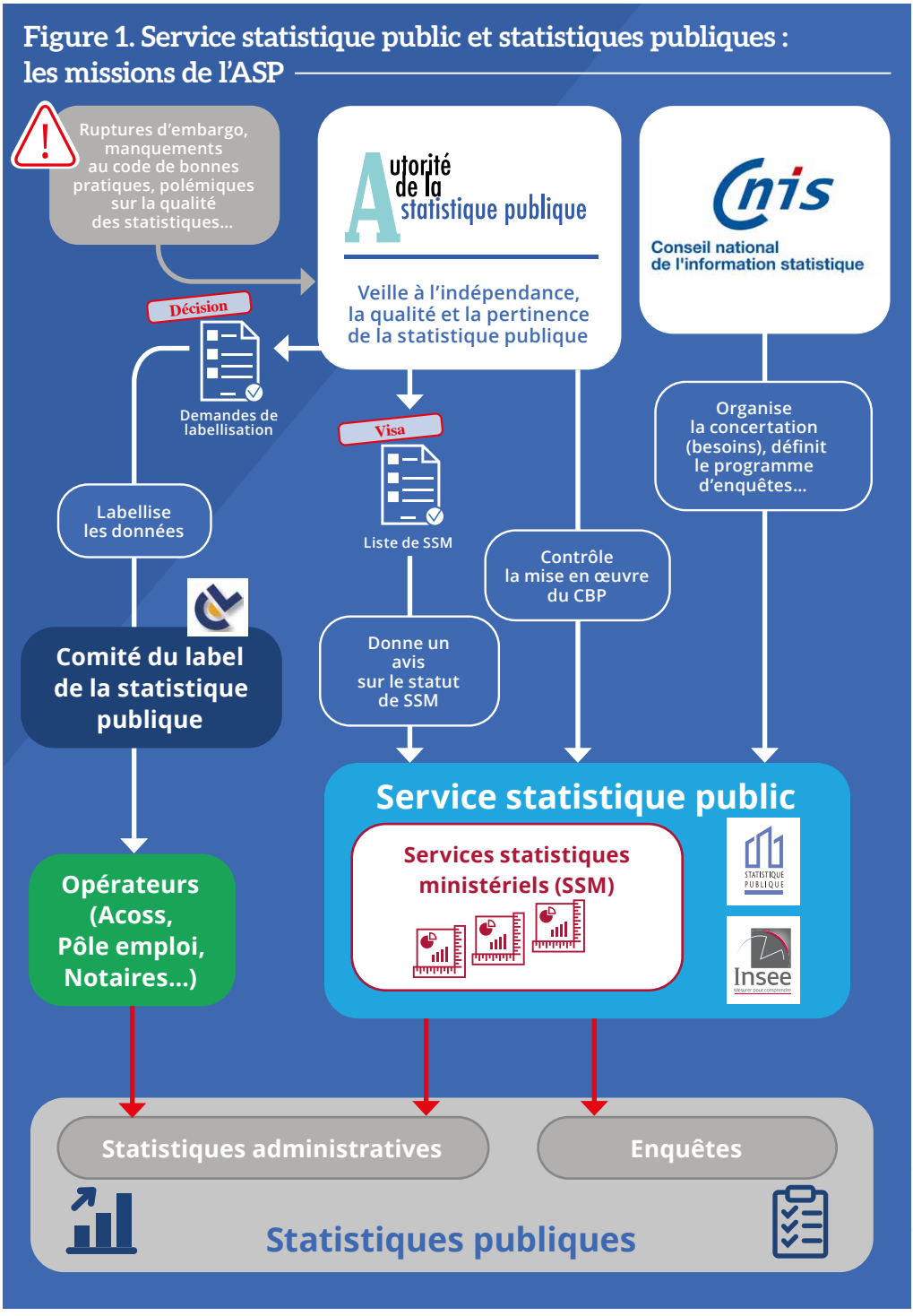
A contrario, le caractère exceptionnel de l'ASP est à relativiser dans la mesure où, dans différents pays, les directeurs généraux des instituts statistiques sont assistés de conseils ayant des missions proches. Toutefois, pour en apprécier le rôle effectif, il est nécessaire d'examiner précisément leurs missions, en distinguant : selon que l'institut statistique est plutôt une instance de synthèse ou un producteur direct de données ; selon que ceux-ci sont focalisés strictement sur l'indépendance professionnelle ou ont un champ d'investigation plus large ; et selon que leur rôle est principalement de contrôle, de prescription ou de conseil.

> *« Avec le recul, les voix qui soulignaient que, dans les organisations, la culture et les capacités comptent autant pour la performance que l'architecture apparente se sont avérées clairvoyantes. »*

Au moment où se discutait la création de l'ASP, une Autorité avait été mise en place au Royaume-Uni, où l'institut statistique[7] est une « agence ». Certes, il était objecté à ce schéma qu'il fallait relativiser l'indépendance d'entités dont le budget reste très dépendant du pouvoir politique et que l'indépendance n'était pas une fin en soi. Cependant, l'idée que ceci serait l'organisation cible « optimale », vers laquelle il faudrait tendre, était largement répandue. Avec le recul, les voix qui soulignaient que, dans les organisations, la culture et les capacités comptent autant pour la performance que l'architecture apparente se sont avérées clairvoyantes ; et l'organisation choisie apparaît pragmatique, préservant l'acquis du service statistique public français, intégré dans l'administration et attaché à l'excellence.

7. L'ONS *(Office for National Statistics).*

Figure 1. Service statistique public et statistiques publiques : les missions de l'ASP



Notre système statistique progresse et, si sa situation n'est sûrement pas parfaite, elle apparaît enviable à beaucoup d'égards, en premier lieu parce que la statistique publique est produite par des statisticiens de haut niveau. Disposant actuellement, comme l'ASP en avait émis le souhait, d'un contrat d'objectifs et de moyens lui permettant de gérer des projets pluriannuels, l'Insee a en fait les attributs d'une « agence » (Bureau et Naves, 2015). Mais il bénéficie des synergies résultant de son intégration administrative, qui justifie sa capacité d'études et lui permet d'offrir des carrières à ses agents. Sa réactivité dans le contexte de la pandémie de la Covid-19[8], lui a permis de tirer parti de la dématérialisation de ses enquêtes pour continuer à mesurer l'activité économique, de la mise en place des données de caisse pour mesurer le pouvoir d'achat et de l'utilisation de nouvelles données pour mesurer les mouvements de population. Cette réactivité est attribuable à ces capacités internes.

L'AUTORITÉ S'APPUIE SUR LE CODE DE BONNES PRATIQUES DE LA STATISTIQUE

Dans un contexte où la publicité des avis de l'ASP constitue la principale incitation à mettre en œuvre ses recommandations, il importe que ceux-ci soient bien argumentés et qu'il ne puisse leur être objecté des incertitudes quant à l'interprétation des principes ou de leur champ d'application. Sinon, il serait facile d'en contester la légitimité ou en atténuer la portée en créant de la confusion ou en invoquant différents éléments circonstanciels, d'ignorance ou d'absence d'intention. À cet égard, le travail de l'Autorité s'appuie d'abord sur la définition précise de l'indépendance professionnelle donnée par la réglementation européenne ***(encadré 2)***.

Par ailleurs, au niveau national, la révision du décret n°2009-250 en 2018 a été l'occasion d'introduire plus de clarté, en posant que l'Autorité émet tout avis qu'elle estime utile *« pour s'assurer du respect, par le service statistique public, des principes du Code de bonnes pratiques de la statistique (CBP) européenne »*[9]. Le respect de ce Code, qui précise les conditions de l'indépendance professionnelle ainsi que les principes de base pour la qualité statistique (objectivité, égal traitement des utilisateurs, fiabilité, confidentialité, efficience, méthodologies solides, documentation des sources et des résultats…) est donc impératif ***(encadré 3)***.

De plus, son contrôle ne fait pas exagérément appel à l'interprétation, les « indicateurs » qui déclinent ses 16 principes étant précis : par exemple, que la satisfaction des utilisateurs doit être mesurée à intervalles réguliers, ce qui est directement vérifiable ; ou encore que les statistiques provenant de différentes sources doivent être comparées et conciliées. En cas de divergences, l'ASP est donc fondée à demander la réalisation des travaux nécessaires à la compréhension des écarts.

Souvent, plusieurs indicateurs sont à considérer. Ainsi, face à une polémique sur la mesure d'une statistique-clé (pouvoir d'achat, chômage, pauvreté, etc.), l'Autorité considérera, outre la définition de l'indépendance professionnelle, l'indicateur qui établit que les responsables statistiques sont seuls compétents pour décider des méthodes et normes statistiques ; et ceux qui demandent que les méthodologies soient fondées sur des

8. Voir l'article de Jean-Luc Tavernier sur le « Fonctionnement de l'Insee dans la période de confinement » dans ce même numéro.
9. Prévu à l'article 2 du règlement (CE) n° 223/2009 du Parlement européen et du Conseil du 11 mars 2009.

considérations statistiques (absence de biais, par exemple) et conformes aux meilleures normes internationales.

Ces références sont alors incontestables. De plus, en ne limitant pas le contrôle au seul premier principe du CBP concernant l'indépendance, le rôle de l'Autorité n'apparaît pas univoque, ce qui concourt à sa crédibilité. En effet, il est plus facile, par exemple, d'exiger le respect le plus strict des règles d'embargo par ceux qui en bénéficient quand on s'est assuré en amont que les dates et heures de parution des statistiques sont annoncées à l'avance ; et que l'égal accès de tous à la statistique publique apparaît comme un souci permanent de son action. Le CBP établissant un cadre qui reconnaît lui aussi que la protection de l'indépendance professionnelle et l'exigence de qualité doivent aller de pair, l'examen des manquements en référence à celui-ci permet non seulement d'en faciliter la correction mais souvent aussi d'enclencher des progrès plus structurels.

« Ces références sont alors incontestables. De plus, en ne limitant pas le contrôle au seul premier principe du CBP concernant l'indépendance, le rôle de l'Autorité n'apparaît pas univoque, ce qui concourt à sa crédibilité. »

UN TIERS GARANT DE L'INDÉPENDANCE DE LA STATISTIQUE PUBLIQUE

Historiquement, l'indépendance de notre statistique publique s'était construite sur deux piliers :

- la déontologie professionnelle des statisticiens, ceux-ci accordant plus d'importance à l'approbation de la qualité des travaux par leurs pairs qu'à celle de toute autre Autorité, gouvernementale ou parlementaire ;
- et le rôle du Cnis, cadre de dialogue structuré et exigeant avec les utilisateurs.

S'appuyant sur ces deux piliers, la statistique publique s'était forgée une réputation méritée. Cependant, ceci n'empêchait pas les polémiques, ni même que celles-ci prennent parfois un tour rugueux, comme cela fût le cas en 1978 à propos de l'évolution du chômage. Le communiqué du directeur général de l'Insee, affirmant qu'en l'espèce l'Insee ne se trompait pas, pouvait suffire alors pour mettre un terme à toute incertitude sur son indépendance. Aujourd'hui, la confiance générale dans les institutions publiques s'est fissurée. Garante de celle-ci, l'ASP veille donc à l'assurer et à en convaincre. Pour cela, elle se réfère principalement au premier principe du CBP ***(encadré 4)***.

Encadré 2. La définition de l'indépendance professionnelle dans les règlements européens

« Les statistiques doivent être développées, produites et diffusées d'une manière indépendante, notamment en ce qui concerne le choix des techniques, des définitions, des méthodologies et des sources à utiliser, ainsi que le calendrier et le contenu de toutes les formes de diffusion, et ces tâches sont accomplies sans subir aucune pression émanant de groupes politiques, de groupes d'intérêt, d'autorités nationales ou d'autorités de l'Union. »

(règlement européen n°2015/759)

UN ENGAGEMENT ATTENDU SUR UNE PROGRAMMATION ANNONCÉE DES PUBLICATIONS

L'ASP est amenée, par exemple, à intervenir lorsque des publications statistiques sont retardées par rapport à des dates annoncées. Ce fut le cas en 2011 après que le président de la Fédération des conseils de parents d'élèves (FCPE) lui eut signalé que le nombre de publications réalisées au cours des trois premiers trimestres de l'année 2011 par le service statistique de l'Éducation nationale lui paraissait inférieur à ce que l'on pouvait attendre au vu de son programme de travail, suspectant un phénomène de dissimulation. L'action de l'ASP a alors permis que le retard soit résorbé et le programme des travaux et des publications pour l'année 2012 rendu public, l'audition de ce SSM permettant finalement de vérifier que la situation était rétablie.

« Plus généralement, l'affichage public des informations que la statistique publique va diffuser dans les mois, trimestres et années à venir représente un engagement fort vis-à-vis de tous les utilisateurs. »

Plus généralement, l'affichage public des informations que la statistique publique va diffuser dans les mois, trimestres et années à venir représente un engagement fort vis-à-vis de tous les utilisateurs. Dès 2009, l'ASP avait donc souhaité que soit élargie la liste des statistiques du SSP dont le calendrier de publication était annoncé à l'avance, allant au-delà des principales statistiques économiques conjoncturelles. L'objectif était de neutraliser toute intervention possible sur les dates de diffusion, tout report devant être exceptionnel, signalé et justifié. L'Insee y a donné suite en mettant en ligne, à partir de 2013, un calendrier annuel de la statistique publique. Fin 2017, l'Autorité constatait aussi la mise en ligne, par tous les SSM, de leur calendrier prévisionnel de diffusion. De plus, l'ASP a demandé à l'Insee d'effectuer un suivi de cette ponctualité pour chacun des SSM, qui permet de constater depuis un taux moyen de ponctualité d'un peu plus de 90 %, les retards constatés ne remettant pas en cause l'indépendance des SSM.

Tous les utilisateurs doivent avoir accès aux publications statistiques au même moment, et dans les mêmes conditions. Tout accès privilégié préalable à la diffusion qui est accordé à un utilisateur extérieur doit donc être limité, suffisamment justifié, contrôlé et rendu public. Certaines informations économiques sont cependant communiquées sous embargo, notamment aux journalistes et aux cabinets ministériels concernés pour leur permettre de prendre connaissance des indicateurs quelques heures avant leur publication. Depuis la création de l'ASP, huit ruptures d'embargo ont été constatées par l'Autorité dont six d'origine gouvernementale. Au-delà du rappel des règles aux personnes concernées, ceci a conduit l'ASP à demander à l'Insee en 2017 de restreindre la diffusion anticipée des indicateurs conjoncturels pour limiter les risques de fuite et d'établir un document-cadre présentant les règles d'embargo pour l'ensemble du service statistique public.

L'examen de ces différents cas a souvent permis de trouver des règles plus adaptées, pour tenir compte notamment de l'évolution des médias. Ainsi, suite à une rupture d'embargo sur les comptes nationaux trimestriels du 4e trimestre 2018, l'Autorité a approuvé la proposition de l'Insee consistant à aligner les horaires de levée d'embargo de tous les indicateurs qui paraîtraient le même jour.

RÉAGIR AUX CRITIQUES OU AUX UTILISATIONS ABUSIVES DES STATISTIQUES

L'ASP intervient par ailleurs lorsque des contestations sur la fiabilité des chiffres sont susceptibles de porter atteinte à la crédibilité de la statistique auprès du public. Les thématiques récurrentes concernent le pouvoir d'achat, l'indice des prix, le chômage, la pauvreté, l'immigration, domaines dans lesquels l'indépendance de l'Insee est vite incriminée. À cet égard, le CBP établit que, s'il y a lieu, les autorités statistiques *« s'expriment publiquement sur les questions statistiques, y compris sur les critiques et les utilisations abusives des statistiques publiques »*.

Il appartient donc d'abord à l'Insee (ou aux services statistiques ministériels) de répondre. Dans le cas où les polémiques prennent de l'ampleur, l'Autorité agit directement. Ce fut le cas en juin 2011, une vive polémique, fortement relayée par les médias, s'étant déclenchée après des déclarations du ministre de l'Intérieur de l'époque sur les statistiques relatives à l'estimation de l'échec scolaire des enfants d'immigrés. L'Insee a alors fait en sorte que tous les journalistes en recherche d'informations sur le sujet aient accès aux statistiques sur ce thème, et publié un communiqué, expliquant ce qui pouvait en être déduit des parcours scolaires des enfants d'immigrés. Parallèlement, le Président de l'ASP a informé le ministre de l'Intérieur de l'écart qui existait entre l'estimation faite par celui-ci et les ordres de grandeur obtenus à partir des statistiques diffusées.

Encadré 3. Les grands principes du Code de bonnes pratiques de la statistique européenne

Les facteurs institutionnels et organisationnels ont une influence non négligeable sur l'efficacité et la crédibilité d'une autorité statistique qui élabore, produit et diffuse des statistiques européennes. Les principes de base sont l'indépendance professionnelle, la coordination et la coopération, le mandat pour la collecte des données, l'adéquation des ressources, l'engagement sur la qualité, le secret statistique, l'impartialité et l'objectivité.

Pour élaborer, produire et diffuser des statistiques européennes, les autorités statistiques appliquent pleinement les normes, les lignes directrices et les bonnes pratiques européennes et internationales dans leurs processus statistiques, tout en cherchant constamment à innover. La crédibilité des statistiques est renforcée par une réputation de bonne gestion et d'efficacité. Les principes de base en sont une méthodologie solide, des procédures statistiques adaptées, une charge non excessive pour les déclarants et un bon rapport coût-efficacité.

Les statistiques disponibles correspondent aux besoins des utilisateurs. Les statistiques respectent les normes de qualité européennes et répondent aux besoins des institutions européennes, des administrations nationales, des instituts de recherche, des entreprises et du public en général. La qualité des résultats est mesurée par le fait que les statistiques sont pertinentes, exactes, fiables, actuelles, cohérentes, comparables entre les régions et les pays, et faciles d'accès pour les utilisateurs, c'est-à-dire à l'aune des principes régissant les résultats statistiques.

Pour plus de détails, voir (Union européenne, 2017).

DES AVIS QUI DÉBOUCHENT PARFOIS SUR LA CRÉATION D'OUTILS

La résolution des situations de crise peut amener à faire évoluer les outils de la statistique publique. C'est ainsi qu'a été créé le simulateur personnalisé d'inflation ou que se sont enrichies les études sur les dépenses contraintes ou sur la précarité énergétique.

Les transformations peuvent aussi être structurelles comme ce fut le cas pour les statistiques en matière de délinquance. En effet, alors qu'en matière de déficit budgétaire, par exemple, les controverses avec le public ne portent que sur l'opportunité et l'évaluation des mesures à prendre, les polémiques sur la sécurité intérieure étaient particulièrement vives au moment où l'ASP se mettait en place et elles mettaient en cause la mesure même des phénomènes, avec une suspicion récurrente de manipulation des chiffres. L'absence d'indépendance professionnelle qui était associée à la production des chiffres fournis au public en ce domaine est alors apparue intenable, ce qui a conduit à la création d'un service statistique ministériel en 2014[10], bénéficiant de toutes les garanties d'indépendance que ceci implique. Le constat que les polémiques inutiles pouvaient ainsi être évitées a progressivement pris le pas sur les craintes qui demeuraient au sein des services par rapport à cette indépendance.

L'AUTORITÉ INTERVIENT DANS UN CONTEXTE DE DÉFIANCE GÉNÉRALISÉE

La confiance des utilisateurs ne dépend pas seulement de l'indépendance professionnelle, mais aussi de l'objectivité, de l'impartialité, de la pertinence et de la qualité des données produites par la statistique publique. À cet égard, la mise en place de l'ASP est intervenue à un moment où la défiance du public tend à s'étendre aussi à l'expertise et tout ce qui se réfère à des normes scientifiques. Par ailleurs, le numérique bouscule la statistique comme le reste de l'économie.

Dans ce nouveau contexte, les attentes du public sont aussi beaucoup plus diversifiées. Il attend que la statistique publique fournisse des données détaillées au-delà des champs habituels de la démographie, de l'économique et du social, sur la délinquance ou le développement durable par exemple. De plus, il a accès, dans tous les domaines, à de nombreuses données alternatives à celles produites par la statistique publique, ceci obligeant ce dernier à convaincre de leur pertinence ou valeur ajoutée, même quand ces alternatives sont très fragiles ou insuffisamment précises.

LABELLISER POUR AUGMENTER LA QUALITÉ ET LE CHAMP DES STATISTIQUES PUBLIQUES

L'ASP est directement impliquée dans le processus d'enrichissement des statistiques publiques puisqu'il lui revient de labelliser comme telles certaines données émanant de l'exploitation des sources produites par les administrations et organismes publics ou privés dans le cadre de missions de service public. Dès 2010, une procédure a été mise en place qui répond à deux objectifs : garantir la qualité des données correspondantes, et accroître le champ des statistiques à même de contribuer au débat public.

10. Le SSMSI, service statistique ministériel de la Sécurité intérieure.

Leur finalité est bien d'améliorer l'information des utilisateurs et répondre à leurs demandes dans un contexte où elles sont de plus en plus diversifiées.

> *« Cette labellisation porte sur des données statistiques spécifiques, et non sur la production statistique de l'organisme dans son ensemble. »*

Cette labellisation porte sur des données statistiques spécifiques, et non sur la production statistique de l'organisme dans son ensemble. Pour les producteurs dont les données sont labellisées, la labellisation constitue un choix stratégique qui impose cependant une indépendance professionnelle dans leur élaboration, une démarche d'amélioration continue des processus de production des statistiques et une politique de diffusion tenant compte des besoins des utilisateurs.

Depuis sa création, 14 labellisations ont été accordées par l'Autorité pour une durée déterminée, généralement de cinq ans. Les avis de labellisation de l'Autorité sont publiés au journal officiel. Ils sont le plus souvent assortis de recommandations à mettre en œuvre, dont le suivi est assuré par l'Autorité, car il en conditionne le renouvellement. Les champs couverts concernent le domaine social (vieillesse et maladie), l'emploi et le chômage, la sécurité routière et le prix des logements.

Encadré 4. Indicateurs de l'indépendance professionnelle dans le Code européen

1.1 : L'indépendance [...] à l'égard des interventions politiques et autres interférences externes dans l'élaboration, la production et la diffusion des statistiques est inscrite dans le droit [...]

1.2 : Les responsables [...] statistiques ont un rang hiérarchique suffisamment élevé pour leur permettre d'avoir des contacts à haut niveau au sein des administrations et organismes publics. Leur profil professionnel est du plus haut niveau.

1.3 : Il appartient aux responsables [...] statistiques de veiller à ce que les statistiques soient élaborées, produites et diffusées en toute indépendance.

1.4 : Les responsables [...] statistiques sont les seuls compétents pour décider des méthodes, des normes et des procédures statistiques ainsi que du contenu et de la date de diffusion des publications statistiques.

1.5 : Les programmes de travail sont publiés et font l'objet de rapports réguliers sur les progrès accomplis.

1.6 : Les publications statistiques sont clairement distinguées des communiqués politiques et diffusées séparément.

1.7 : S'il y a lieu, [les responsables statistiques] s'expriment publiquement sur les questions statistiques, y compris sur les critiques et les utilisations abusives des statistiques.

1.8 : Les procédures de recrutement et de nomination des responsables [...] statistiques sont transparentes et exclusivement fondées sur des critères professionnels. Les motifs de fin de fonctions sont fixés par le cadre juridique. Il ne peut s'agir de raisons susceptibles de mettre en péril l'indépendance professionnelle ou scientifique.

L'EXEMPLE DES DONNÉES CONCERNANT LE MARCHÉ DU TRAVAIL

En 2014, l'Autorité a ainsi labellisé les statistiques mensuelles des demandeurs d'emploi en fin de mois (DEFM) inscrits à Pôle emploi. Cette labellisation avait été assortie de recommandations, notamment que le commentaire privilégie la tendance des derniers mois et que la faible signification de la variation d'un mois sur l'autre en dessous d'un certain seuil soit mentionnée dans la publication, alors mensuelle. Compte tenu de l'importance prise dans le débat public par ce sujet, les modalités de mise en œuvre de ces recommandations ont été réévaluées en 2015. Finalement, un consensus s'est dessiné pour une trimestrialisation de la publication, car il apparaissait que le commentaire des chiffres mensuels, sur lesquels se portait largement l'attention des observateurs, était sans pertinence statistique en général. De cette manière, les chiffres des DEFM sont désormais publiés avec la même périodicité que l'autre thermomètre du chômage, le taux de chômage de l'Insee, défini au sens du Bureau International du Travail (BIT).

> *« Finalement, un consensus s'est dessiné pour une trimestrialisation de la publication, car il apparaissait que le commentaire des chiffres mensuels, sur lesquels se portait largement l'attention des observateurs, était sans pertinence statistique en général. »*

À cet égard, pour que la diversité des sources ne soit pas un facteur d'incertitude pour le public, l'Autorité a demandé aussi que soient menés des travaux pour expliquer les écarts entre les évolutions des deux indicateurs. Les travaux d'appariement entre le fichier historique des DEFM et l'enquête Emploi menés à cette fin ont permis de constater que les écarts persistants ou de grande ampleur entre les deux séries tiennent aux écarts de concepts pour la mesure du chômage, tels que la disponibilité pour prendre un emploi, par exemple, renvoyant donc à la diversité des situations au sein du « halo » du chômage.

Dans le même esprit, l'ASP a demandé que soient examinés les écarts entre les différentes sources concernant la mesure de l'emploi, ce qui a conduit à un plan d'action pour corriger le biais identifié sur l'emploi des jeunes dans l'enquête Emploi et pour renforcer la communication sur les sources statistiques concernant l'emploi.

Quoique certains aspects nécessitent encore d'être mieux compris, notamment concernant la volatilité des DEFM ou certains biais dans l'évaluation de l'emploi dans les enquêtes par exemple, l'utilisation qui peut être faite des différentes sources pour poser le diagnostic sur le marché du travail a ainsi progressé.

UNE MISSION ÉLARGIE POUR RÉPONDRE AUX NOUVEAUX DÉFIS

Ce bilan met en exergue que la création de l'Autorité de la statistique publique a conduit à une gouvernance équilibrée et que l'extension de sa mission au-delà de ce qui concerne strictement l'indépendance professionnelle s'est révélée très fructueuse, dans un contexte de bouleversement des données auxquelles a accès le public et des attentes de celui-ci.

Les fausses informations et les interprétations erronées des faits et des statistiques sont plus répandues que jamais dans le discours public et pénètrent souvent la sphère politique. Les médias classiques ont une audience de plus en plus faible au profit des réseaux sociaux, sur lesquels les démentis sont partagés six fois moins que les informations fausses. Ceci impose aux autorités statistiques des responsabilités nouvelles et exigeantes. Tous les acteurs de la statistique publique sont concernés pour relever ces défis. L'ASP y concourt en veillant au respect de l'ensemble des principes du Code de bonnes pratiques de la statistique.

« Les médias classiques ont une audience de plus en plus faible au profit des réseaux sociaux, sur lesquels les démentis sont partagés six fois moins que les informations fausses. »

BIBLIOGRAPHIE

ASP, 2020. *Autorité de la statistique publique.* [en ligne]. [Consulté le 20 novembre 2020]. Disponible à l'adresse : *https://www.autorite-statistique-publique.fr/*.

BUREAU, Dominique et NAVES, Marie-Cécile, 2015. *Quelle action publique pour demain ? Cinq objectifs, cinq leviers*. [en ligne]. Avril 2015. Rapport de France Stratégie. [Consulté le 20 novembre 2020]. Disponible à l'adresse : *https://www.strategie.gouv.fr/sites/strategie.gouv.fr/files/atoms/files/rapport_action_publique_bat_13042015.pdf*.

CHAMPSAUR, Paul, 2009. L'Autorité de la statistique publique. In : *Courrier des statistiques.* [en ligne]. Septembre-décembre 2009. N°128, pp. 5-8. [Consulté le 20 novembre 2020]. Disponible à l'adresse : *https://www.epsilon.insee.fr/jspui/bitstream/1/122489/1/cs128.pdf*.

INSEE, 2019. Charte des services statistiques ministériels. In : *site de l'Insee.* [en ligne]. 3 juillet 2020. Le service statistique public. [Consulté le 20 novembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/1302192/Charte_SSM_2019.pdf*.

ISNARD, Michel, 2018. Qu'entend-on par statistique(s) publique(s) ? In : *Courrier des statistiques*. [en ligne]. 6 décembre 2018. N°N1. [Consulté le 20 novembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3646978/courstat-1-10.pdf*.

LECLAIR, Marie, 2019. Utiliser les données de caisse pour le calcul de l'indice des prix à la consommation. In : *Courrier des statistiques.* [en ligne]. 19 décembre 2019. N°N3 , pp. 61-75. [Consulté le 20 novembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254225/courstat-3-6.pdf*.

UNION EUROPÉENNE, 2017. *Code de bonnes pratiques de la statistique européenne.* [en ligne]. 16 novembre 2017. [Consulté le 20 novembre 2020]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/documents/4031688/9332274/KS-02-18-142-FR-N.pdf/130905e7-45a7-4475-b37c-8f699b5e33e1*.

RÉFÉRENCES JURIDIQUES

Décret n° 2009-250 du 3 mars 2009 relatif à l'Autorité de la statistique publique. In : *site de Légifrance.* [en ligne]. Modifié le 20 septembre 2018. [Consulté le 20 novembre 2020]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/LEGITEXT000020344133/2018-09-23/*.

Loi n° 2008-776 du 4 août 2008 de modernisation de l'économie. In : *site de Légifrance.* [en ligne]. Modifiée le 5 juillet 2019. [Consulté le 20 novembre 2020]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000019283050/2020-11-20/*.

Loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques. In : *site de Légifrance.* [en ligne]. Modifié le 28 juin 2010. [Consulté le 20 novembre 2020]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573/2020-11-20/*.

Règlement (CE) n° 223/2009 du Parlement européen et du Conseil du 11 mars 2009 relatif aux statistiques européennes. In : *Journal officiel de l'Union européenne.* [en ligne]. Modifié le 25 avril 2015. [Consulté le 20 novembre 2020]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:02009R0223-20150608*.

Règlement (UE) 2015/759 du Parlement européen et du Conseil du 29 avril 2015 modifiant le règlement (CE) n° 223/2009 relatif aux statistiques européennes. In : *Journal officiel de l'Union européenne.* [en ligne]. [Consulté le 20 novembre 2020]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32015R0759&qid=1605887489757&from=FR*.

LE COMITÉ DU LABEL

UN ACTEUR DE LA GOUVERNANCE AU SERVICE DE LA QUALITÉ DES STATISTIQUES PUBLIQUES

Marc Christine[*] *et Nicole Roth*[**]

Le Comité du label a été créé en 1994 pour attester de la qualité des enquêtes de la statistique publique. Depuis les années deux-mille-dix, son positionnement et ses missions ont évolué, la modification de la loi de 1951 ayant posé une définition élargie des « statistiques publiques ». Le Comité reste sollicité pour instruire « en conformité » les enquêtes ayant reçu un avis d'opportunité du Conseil national de l'information statistique, définissant le programme des enquêtes ayant le visa ministériel. Il intervient désormais aussi à la demande de l'Autorité de la statistique publique, lors de la labellisation de statistiques administratives, produites par des organismes chargés d'une mission de service public ne faisant pas partie du Service statistique public (SSP).

Le Comité a construit au fil du temps une méthode et une jurisprudence pour examiner les dossiers qui lui sont soumis. Partant de considérations exprimées en termes de charge ou de proportionnalité de la collecte aux objectifs poursuivis, le Comité a élargi ses règles d'examen pour couvrir l'ensemble des dimensions de la qualité statistique, telles que formalisées dans le Code de bonnes pratiques de la statistique européenne. Pour le SSP, le Comité du label constitue ainsi un levier pour s'assurer du respect de ces principes, que ce soit en termes de concertation, de qualité méthodologique, de charge proportionnée, de diffusion ou de mise à disposition des sources statistiques dûment documentées.

The Official Statistics Label Committee was set up in 1994 to certify the quality of statistical surveys. Since the 2010s, its missions have extended, as the amendment of the 1951 Statistical Act laid down a broader definition of "official statistics". The Committee continues to assess the quality of surveys after the opportunity statement delivered by the National Council for Statistical Information; it delivers a "conformity" statement, leading to the definition of the program of surveys with a ministerial visa. The Committee now also intervenes at the request of the Official Statistics Authority, in the process of labelling administrative statistics produced by bodies entrusted with a public service mission but being outside the Official Statistical System.

The Committee has built up a method and a body of rules for examining the cases submitted to its examination. Starting from considerations expressed in terms of burden or proportionality of the collection to the objectives pursued, the Committee has extended its rules to cover all dimensions of statistical quality, as formalised in the European Statistics Code of Practice. For the Official Statistical System, the Label Committee thus constitutes a lever for ensuring compliance with the principles of the Code, such as large consultation, methodological quality, proportionate burden, dissemination of databases and availability of metadata.

* À la date de la rédaction de l'article, rapporteur du Comité du label de la statistique publique,
marc.christine@insee.fr
** Présidente du Comité du label de la statistique publique,
nicole.roth@insee.fr

Depuis sa création en 1994 au sein du Conseil national de l'information statistique (Cnis), le Comité du label joue un rôle décisif dans l'élaboration du programme des enquêtes de la statistique publique en France. Bien loin d'être une chambre d'enregistrement des projets d'enquêtes, le Comité atteste de leur « conformité » à des standards de qualité et propose aussi des voies d'amélioration. Compte tenu de son large spectre d'intervention, il joue ainsi un rôle pédagogique, de conseil, voire d'émetteur d'alerte et contribue à la diffusion des bonnes pratiques ainsi qu'à leur documentation.

Dans les années deux-mille-dix, les missions du Comité du label se sont élargies : la labellisation concerne désormais également les statistiques issues de l'exploitation de données administratives, dès lors que celles-ci sont produites par des organismes ayant des missions de service public. Le Comité exerce cette mission pour le compte de l'Autorité de la statistique publique (ASP)[1], qui prononce les avis officiels. L'évolution des missions du Comité a conduit à modifier son appellation en **Comité du label de la statistique publique**[2] pour marquer son nouveau positionnement.

Le Comité du label s'inscrit depuis ses débuts dans une démarche contribuant à la qualité des opérations statistiques, en parallèle de la dynamique initiée au niveau européen. Il a ainsi accompagné et contribué depuis le début des années quatre-vingt-dix à promouvoir les préoccupations de qualité statistique, selon une jurisprudence qui s'est construite progressivement.

DU COMITÉ DU LABEL AU SEIN DU CNIS…

Le Comité a été créé en 1994 à la demande de l'assemblée plénière du Conseil national de l'information statistique (Cnis), sous la dénomination de **Comité du label des enquêtes statistiques des services publics et des autres services producteurs d'informations statistiques**. Positionné au sein du Cnis, il avait pour mission *« d'examiner la conformité des projets d'enquêtes statistiques après que celui-ci en ait reconnu l'opportunité et de proposer un label d'intérêt général ; de proposer aux ministres compétents la délivrance du visa prévu par l'article 2 de la loi de 1951 modifiée »*[3].

Il s'agissait d'améliorer les procédures d'élaboration des programmes statistiques par une attention particulière portée aux projets d'enquêtes (Allain, 1995). L'analyse ayant conduit à cette initiative était qu'il fallait consolider le dispositif du Cnis qui ne répondait plus suffisamment à certaines exigences qui jouaient de façon croissante. Pour les entreprises, notamment les plus petites, il s'agissait de faire accepter la charge que constitue pour elles la réponse aux enquêtes statistiques, de distinguer les enquêtes statistiques des formalités administratives et de démontrer leur finalité et leur utilité. Pour les ménages, il s'agissait de se prémunir

> *« La phase d'expérimentation a permis de définir de façon heuristique les premières règles de cet examen de conformité. »*

1. Voir l'article de Dominique Bureau sur « L'Autorité de la statistique publique : dix ans d'activité, pour une statistique indépendante et de qualité », dans ce même numéro.
2. Par commodité, dans la suite de l'article, le nom officiel sera remplacé par Comité du label, ou plus simplement encore par Comité.
3. Voir également (Isnard, 2018) pour ce qui concerne la définition des statistiques publiques.

de réactions des enquêtés ou de plaintes auprès de la Cnil[4] quant au volume et à la nature des questions posées.

Le nouveau Comité du label a été créé au sein du Cnis pour une durée expérimentale de deux ans d'abord, avant d'être pérennisé. La phase d'expérimentation a permis de définir de façon heuristique les premières règles de cet examen de conformité : respect des nomenclatures, cohérence des unités statistiques, prise en compte dans l'échantillonnage de la nécessité de limiter la charge des enquêtés, justification du caractère plus ou moins utile ou intrusif de certaines questions relativement aux objectifs poursuivis, respect de la confidentialité des données individuelles, modes de diffusion des résultats.

… AU COMITÉ DU LABEL DE LA STATISTIQUE PUBLIQUE

La loi de Modernisation de l'économie du 4 août 2008, qui a conduit à la création de l'Autorité de la statistique publique, a aussi porté une définition extensive des statistiques publiques, inscrite dans la loi fondatrice n°51-711 du 7 juin 1951 (modifiée) sur l'obligation, la coordination et le secret en matière de statistiques. La définition des statistiques publiques (article 2 de la loi) regroupe *« l'ensemble des productions issues :*

- *des enquêtes statistiques dont la liste est arrêtée chaque année par un arrêté du ministre chargé de l'Économie ;*
- *de l'exploitation, à des fins d'information générale, de données collectées par des administrations, des organismes publics ou des organismes privés chargés d'une mission de service public.*

La conception, la production et la diffusion des statistiques publiques sont effectuées en toute indépendance professionnelle ».

Au champ des enquêtes dont le programme est arrêté chaque année (et qui ont donc un visa ministériel), s'est rajouté le champ de statistiques diffusées par des opérateurs publics (champ qualifié ici de « péri-SSP »). L'ASP a la charge de labelliser certaines statistiques diffusées par des opérateurs publics à partir de données administratives[5], ce qu'elle fait en s'appuyant depuis 2013 sur l'instruction assurée par le Comité du label, renommé **Comité du label de la statistique publique.** Celui-ci a désormais en charge trois missions (Christine, 2016) ***(encadré 1)***, car à la mission historique d'examen de la conformité des enquêtes de la statistique publique se sont rajoutées :

- **une mission pour le compte de l'ASP**, sur le champ des opérateurs publics lorsqu'ils diffusent des données à des fins d'information générale ;
- **une mission pour le compte du Cnis**, pour juger de la qualité des statistiques produites par des organismes de droit privé. Ceci fait suite aux propositions d'un rapport du Cnis de 2010, sur les statistiques du logement (Vorms, Jacquot et Lhéritier, 2010). La mission s'exerce sur la base du volontariat, contrairement à la labellisation par l'ASP ; en effet, il n'y a pas de support légal encadrant la diffusion des statistiques produites par des organismes de droit privé n'exerçant pas de mission de service public.

Avec cet élargissement de ses compétences, le Comité du label est devenu en 2013 une entité à part. Ses moyens sont assurés par l'Insee, au sein de la direction de la Méthodologie et de la coordination statistique et internationale (DMCSI), mobilisant un rapporteur

4. Commission nationale Informatique et libertés.
5. Produites par exemple par l'Agence centrale des organismes de sécurité sociale (Acoss), Pôle emploi, la Caisse nationale d'allocations familiales (Cnaf), le Service des retraites de l'État (SRE), le Centre d'épidémiologie sur les causes médicales de décès (CépiDC) ou l'Observatoire national interministériel de la sécurité routière (ONISR).

et un secrétariat général, ainsi que des experts pour assurer l'instruction des dossiers. Les enquêtes et les statistiques administratives sont examinées en commissions thématiques, dont la composition est fixée par arrêté, sous la responsabilité du président du Comité, nommé par arrêté du ministre en charge de l'Économie, après avis des présidents de l'ASP et du Cnis[6]).

UN STATUT D'ENQUÊTE DE LA STATISTIQUE PUBLIQUE QUI APPORTE DES GARANTIES

Le statut d'enquête de la statistique publique génère pour les services des contraintes (matérialisées par la procédure même du Cnis et du Comité du label, avec l'exigence de présentation de documents descriptifs détaillés), mais il confère aussi des droits pour les services et apporte des garanties aux répondants.

Les services producteurs d'enquêtes peuvent ainsi mettre en avant la marque « statistique publique » et le critère d'intérêt général, pour se démarquer dans leur communication auprès des enquêtés de la multitude d'enquêtes ou de sondages réalisés à d'autres fins. Ils peuvent aussi demander l'octroi du caractère obligatoire[7], qui constitue un levier pour favoriser la réponse, même si, en pratique, les services cherchent d'abord à persuader les enquêtés de l'intérêt de répondre, en s'appuyant sur la reconnaissance de l'enquête par le Cnis (Le Gléau *et alii*, 2016)..

Enfin, le statut d'enquête de la statistique publique engage le producteur à prendre toutes les mesures pour assurer le secret statistique lors de la diffusion des résultats et à ne pas communiquer les données individuelles recueillies à des fins de contrôles. Il permet d'apporter aux personnes ou entités qui fournissent des informations utilisées pour l'établissement de statistiques l'assurance que ces informations ne seront pas utilisées d'une façon susceptible de leur porter préjudice. Cette garantie concerne la protection de la vie privée, du secret commercial ou des affaires, mais aussi dans certains cas les administrations (par exemple les établissements scolaires ou sociaux). Ainsi, dans les domaines de la protection sociale ou de l'éducation, par exemple, certaines enquêtes menées auprès de services administratifs ne nécessitent pas théoriquement un statut d'enquête de la statistique publique, car leur collecte repose sur des personnels faisant partie de l'administration. Leur conférer ce statut d'enquête leur permet cependant de se démarquer de logiques de pure gestion administrative relevant notamment du Code des relations entre le public et les administrations (CRPA).

> *« Les services producteurs d'enquêtes peuvent ainsi mettre en avant la marque « statistique publique ». »*

6. L'arrêté du 2 mai 2013 définit les modalités d'organisation du Comité du label.
7. Le défaut de réponse à une enquête obligatoire peut donner lieu à une décision d'amende administrative, voire d'une amende pénale, prise par le ministre chargé de l'Économie, après avis du Comité du contentieux des enquêtes statistiques obligatoires.

LE CHEMINEMENT DES ENQUÊTES DE LA STATISTIQUE PUBLIQUE

Le périmètre des enquêtes de la statistique publique (au sens de la loi de 1951) se définit de façon quasi tautologique : est qualifiée « d'enquête de la statistique publique », et dotée d'un visa du ministère de l'Économie, toute enquête publiée dans l'arrêté annuel (ou ses avenants) qui définit le programme d'enquêtes.

Trois critères sont *a priori* nécessaires pour intégrer ce périmètre :

- un critère organique : l'enquête doit être réalisée par un service public ou assimilé ;
- un critère de finalité : l'obtention de statistiques ;
- et un critère de champ d'application : l'enquête nécessite le concours de personnes étrangères à l'administration.

Encadré 1. Définition réglementaire des missions du Comité du label

(Extrait du décret n° 2009-318 relatif au Conseil national de l'information statistique, au comité du secret statistique et au comité du label de la statistique publique)

Article 20 :

« I – Le comité du label examine pour le compte du Conseil national de l'information statistique les projets comportant la collecte d'informations au moyen d'enquêtes pour lesquelles est sollicité le visa prévu à l'article 2 de la loi du 7 juin 1951 susvisée. [...]

Il vérifie que ces projets :

a) ont reçu un avis d'opportunité favorable d'un président d'une commission thématique du Conseil national de l'information statistique ; ou

b) sont prévus par une loi spéciale ; ou

c) présentent un caractère de nécessité et d'urgence indiscutables.

Il évalue les modalités de mise en œuvre prévues par le service producteur, notamment en prenant en compte la qualité statistique du projet, la charge qu'implique l'enquête pour les personnes physiques ou morales qui en font l'objet, le degré de concertation avec les utilisateurs et le respect des termes de l'avis d'opportunité. En cas d'évaluation favorable du projet, il donne à l'enquête un avis de conformité ainsi qu'un avis sur son caractère obligatoire.

Le comité examine pour le compte du Conseil national de l'information statistique et à la demande de ce dernier les statistiques produites par des organismes de droit privé.

Il évalue notamment la qualité statistique des processus ayant conduit à ces statistiques et leur conformité aux règles de l'art reconnues par la profession. Il transmet les conclusions de cet examen au président du Conseil national de l'information statistique.

II. – Le comité du label de la statistique publique examine pour le compte de l'Autorité de la statistique publique et à la demande de celle-ci les processus d'exploitation et de diffusion, à des fins d'information générale, de données collectées par des administrations, des organismes publics et des organismes privés chargés d'une mission de service public. Le résultat de cet examen est traduit dans un avis.

Ces critères sont assez généraux. *In fine*, c'est la procédure elle-même qui circonscrit *a posteriori* le périmètre des enquêtes de la statistique publique. En pratique, le visa du ministre, avec mention éventuelle de l'obligation de réponse, est octroyé, par délégation, par le directeur général de l'Insee.

Le cheminement des projets d'enquêtes pour obtenir le visa est le suivant ***(figure 1)*** : ils doivent en premier lieu obtenir un **avis d'opportunité** positif en commission thématique du Cnis, qui atteste que l'enquête est utile, qu'elle répond à un besoin d'intérêt général et qu'elle ne fait pas double emploi avec les sources existant sur le même sujet. Pour les enquêtes régionales, c'est le comité régional pour l'information économique et sociale (Cries) ou, en son absence, une instance régionale dûment constituée et réunie par le directeur régional de l'Insee pour l'occasion, qui se prononce en premier sur l'opportunité de l'enquête, avant de transmettre son avis au Cnis.

Deux cas d'urgence sont cependant prévus, soit dans le cadre d'une loi spéciale, soit en raison d'un caractère de nécessité et d'urgence indiscutables. Il est alors admis que les enquêtes sont opportunes par nature, et donc dispensées, si les délais sont très tendus, d'un examen en commission thématique du Cnis. Cette modalité a été utilisée en 2020 notamment pour les enquêtes visant à assurer un suivi économique ou social des effets de la crise sanitaire liée à la Covid-19 (enquêtes Acemo-Covid de la Dares et EpiCov de la Drees et de l'Inserm)[8].

La seconde étape consiste à obtenir un **avis de conformité** du Comité du label, sur la base d'une instruction très complète analysant le processus prévu pour réaliser l'enquête, depuis sa conception jusqu'à la diffusion des résultats.

L'enquête bénéficie alors d'un label d'intérêt général et de qualité statistique délivré par le Cnis, assorti le cas échéant d'une proposition de rendre l'enquête obligatoire pour les répondants. En pratique, par délégation formelle du Cnis, le label est attribué par le président du Comité du label, dès lors qu'il délivre son avis de conformité.

INTÉRÊT GÉNÉRAL ET QUALITÉ STATISTIQUE : UNE JURISPRUDENCE ÉTABLIE AU FIL DU TEMPS…

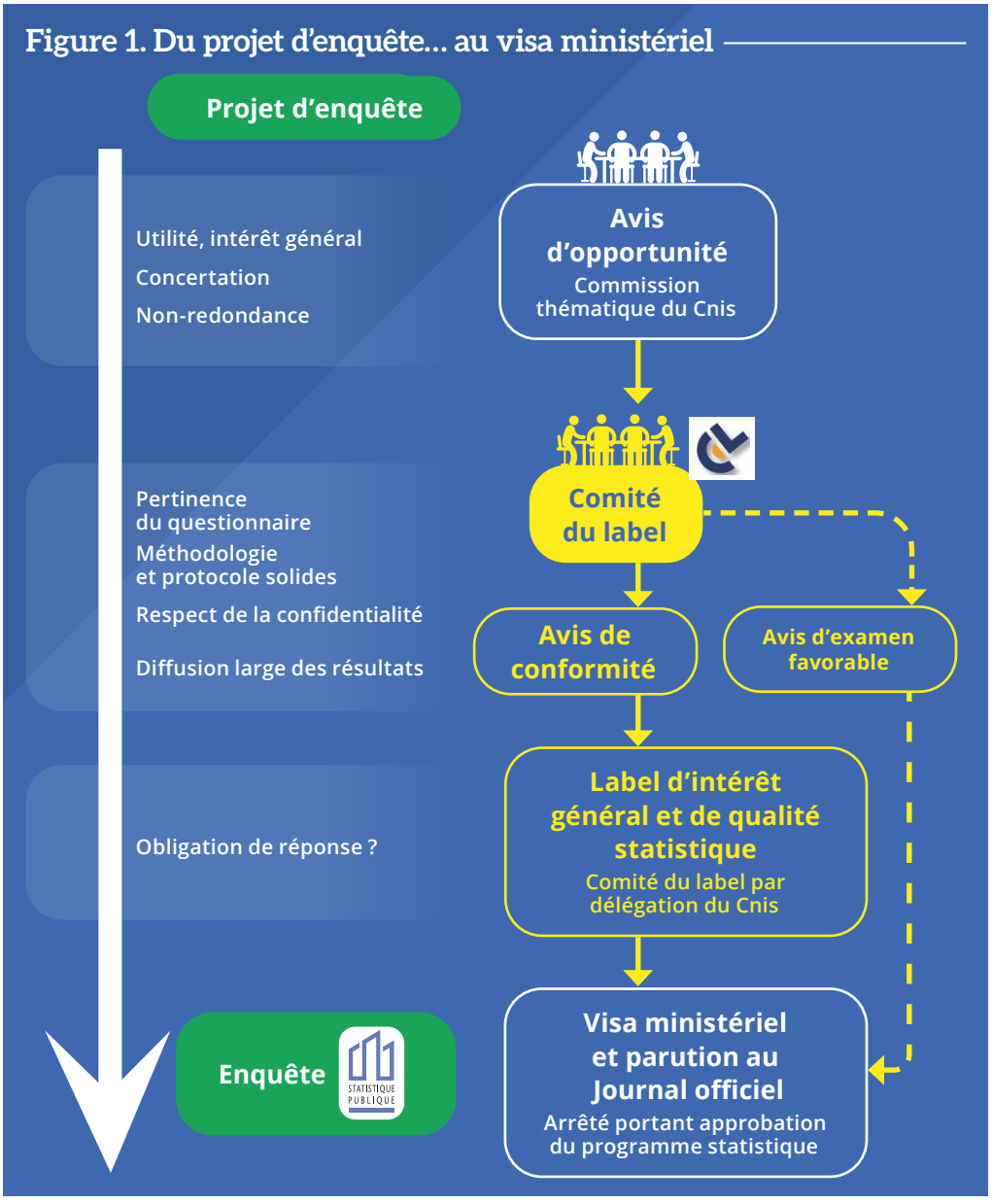
Les textes réglementaires laissent au Comité du label toute latitude pour apprécier la conformité des enquêtes et la qualité statistique. Depuis 25 ans, le Comité a bâti au fil du temps une jurisprudence définissant les critères attendus pour justifier de l'intérêt général et de la qualité de l'enquête, notamment :

- la *concertation* avec les représentants des utilisateurs potentiels des résultats de l'enquête et des organisations syndicales ou professionnelles concernées ;
- la *teneur du questionnaire* et le *caractère non excessif des questions posées,* ainsi que leur conformité aux objectifs entérinés par le Cnis, ce qui conduit à veiller à la non-redondance des informations recueillies par rapport à d'autres sources et la proportionnalité des questions posées au regard des objectifs visés ;

8. Voir l'article de Jean-Luc Tavernier sur le « Fonctionnement de l'Insee en période de confinement » paru dans ce même numéro.

- la *méthodologie de la collecte et des traitements*, qui renvoie à un vaste ensemble de questions sur l'échantillonnage, le protocole et le mode de collecte, les redressements de la non-réponse, les contrôles de qualité prévus tout au long des traitements, afin d'évaluer la précision des indicateurs et leur absence de biais. À ces questions de mesure se rajoutent les questions de définitions des concepts en amont de leur mesure, et le cas échéant, de conformité à des règles émanant de règlements européens ou internationaux. Enfin, le protocole de collecte est aussi examiné de façon à assurer une collecte loyale vis-à-vis des enquêtés sollicités pour répondre, visant à assurer leur information sur les objectifs de la collecte ainsi que sur leurs droits ;

Figure 1. Du projet d'enquête… au visa ministériel



- les mesures prises pour assurer la *confidentialité* et le *respect du secret statistique ;*
- la *diffusion* large des résultats et l'ouverture des données individuelles notamment aux chercheurs, dans le respect de la confidentialité.

Pour apprécier ces critères, le Comité a établi un dossier-type, qui permet aux services demandeurs de documenter de manière normalisée l'ensemble du processus d'enquête défini lors de la phase de conception et prévu pour le traitement des données en aval après la collecte de l'enquête. Le Comité est particulièrement sensible à la phase de concertation en amont (le service doit indiquer et objectiver comment il a consulté les parties prenantes et les organisations syndicales) et à la phase de tests des questionnaires (pour apprécier leur longueur mais aussi la bonne compréhension des questions par les enquêtés). Le Comité veille aussi aux aspects méthodologiques et notamment au dimensionnement de l'échantillon pour atteindre la précision attendue et aux redressements prévus pour s'assurer de l'absence de biais de réponse et corriger les effets négatifs de la non-réponse. En cas d'évolution d'une enquête régulière, il examine en particulier les dispositions prévues pour contrôler d'éventuelles ruptures de séries et de leur documentation auprès des utilisateurs. Enfin, il veille à ce que les moyens prévus soient proportionnés aux objectifs. L'ouverture des données aux chercheurs (dans le respect de la confidentialité) constitue un dernier critère important, les enquêtes de la statistique publique constituant des « biens publics » à partager.

… QUI S'INSCRIT DANS LA DÉMARCHE EUROPÉENNE SUR LA QUALITÉ

Sans y faire explicitement référence à l'origine, ces critères se rapprochent de ceux définis dans le cadre de la statistique officielle européenne, avec l'attention portée à la démarche qui s'est fait jour depuis les années quatre-vingt-dix. Dès 2001, six critères de qualité ont ainsi été énoncés par Eurostat et officiellement retenus par le Comité du programme statistique européen : pertinence, précision, actualité, comparabilité et cohérence (Desrosières, 2003). Ils ont été consolidés depuis avec l'élaboration du Code de bonnes pratiques de la statistique européenne et de ses 16 principes (Union européenne, 2017).

Les démarches de labellisation impulsées dans le cadre du Cnis et du Comité du label s'inscrivent de fait dans cette logique. Elles ont d'ailleurs été identifiées par les auditeurs européens dans le cadre de la revue par les pairs, réalisée en 2014, comme des éléments de démonstration de la qualité des enquêtes.

> *« Les démarches de labellisation impulsées dans le cadre du Cnis et du Comité du label s'inscrivent de fait dans cette logique. »*

Ainsi, la concertation initiée par le Cnis constitue un moyen pour assurer la pertinence des enquêtes, comme répondant aux besoins exprimés par les utilisateurs. L'examen en conformité permet quant à lui de s'assurer que la conception des enquêtes est conforme aux règles de l'art en termes méthodologiques, mais aussi déontologiques. Enfin, les dossiers constitués pour le Comité du label constituent bien souvent un élément important de la documentation des enquêtes de la statistique publique (les « métadonnées ») qui permet d'assurer leur transparence pour les utilisateurs (Bonnans, 2019).

QUELQUES CAS D'ENQUÊTES DONT LA QUALITÉ N'EST PAS ATTESTÉE *A PRIORI*

Il peut arriver que le Comité du label constate que l'enquête qui lui est présentée est bien d'intérêt général, mais qu'elle ne peut pas *ex ante* être considérée comme ayant les qualités suffisantes pour recevoir le label plein et entier, parce qu'elle ne satisfait pas d'emblée à certains des critères énoncés ci-dessus. Le Comité du label peut cependant émettre un **avis d'examen favorable** et demander, le cas échéant, l'octroi du visa ministériel ***(figure 1)***.

Ces cas particuliers relèvent notamment des trois situations suivantes :

- **enquêtes expérimentales ou pilotes** : il s'agit d'enquêtes mettant en œuvre des techniques d'échantillonnage ou des modes de questionnement présentant un caractère novateur marqué ou dont les résultats strictement méthodologiques ne sont pas destinés à être publiés en tant que tels (par exemple, lors de la rénovation de l'enquête Emploi, ou pour instruire la question d'éventuels effets de mode de collecte, en face-à-face, téléphone ou par internet). Le caractère expérimental de ces travaux conduit le plus souvent le Comité du label à en prendre acte sans attribuer le label, car il n'est pas en mesure de se prononcer *a priori* sur la conformité de méthodes, précisément en raison de leur caractère novateur ; la diffusion large des résultats de ces enquêtes n'est pas non plus nécessairement assurée, précisément en raison de leur caractère exploratoire ;
- **enquêtes participant à l'évaluation de politiques publiques** : il s'agit d'enquêtes mobilisant des méthodologies de modélisation complexes sur lesquelles il est délicat de se prononcer *ex ante* (par exemple, concernant l'enquête sur l'expérimentation des mesures « zéro chômeur de longue durée dans les territoires ») ; en raison de cette incertitude sur la qualité ou la définition des méthodes aval et aussi pour éviter un risque d'instrumentalisation, le Comité évite alors de se prononcer sur la qualité *a priori* des résultats de ces enquêtes, cette responsabilité relevant alors d'autres instances, de type conseil scientifique ;
- **enquêtes qui font l'usage de protocoles de mesure particuliers**, incluant par exemple des mesures biologiques ou médicales ou des tests techniques (auprès des personnes ou auprès des logements : capteurs de poussières, données énergétiques ou sur la qualité de l'air, etc.), qui ne peuvent être validés par des non-spécialistes de ces domaines. Le Comité du label peut alors ne donner son avis que sur une partie restreinte du processus de l'enquête.

UN AVIS SIMPLE POUR DES ENQUÊTES REQUÉRANT UNIQUEMENT UN ÉCHANTILLON DE L'INSEE

Certaines enquêtes réalisées à des fins de recherche scientifique ou historique, ou dans le cadre de recherches-action (par exemple sur la connaissance des mobilités quotidiennes au niveau de certains territoires) peuvent sous certaines conditions obtenir un échantillon probabiliste issu du Recensement de la population ou du Fichier démographique sur les logements et les personnes issus de sources fiscales (Fideli).

Ces enquêtes sont souvent réalisées dans le cadre d'infrastructures européennes dotées de leurs propres instances de gouvernance (c'est le cas par exemple de l'Enquête Sociale Européenne *(European Social Survey)* qui est menée tous les deux ans dans la plupart des pays européens). La difficulté pour l'examen de ces projets réside souvent dans le fait que les recommandations du Comité se heurtent alors à d'autres logiques de gouvernance et de spécifications supra-nationales et sont donc généralement peu suivies d'effet, notamment s'agissant du questionnaire ou de certains aspects méthodologiques, définis et entérinés en amont par un consortium international.

> *« L'esprit de l'examen par le Comité du label est depuis sa création de s'assurer de la conformité des méthodes aux règles de l'art, des nuances et des précisions se sont progressivement dessinées dans ses pratiques, conduisant à dresser une typologie de ses avis : avis de conformité, avis d'examen favorable avec visa, avis d'examen simple (sans visa). »*

Le Comité du label est alors sollicité par le Comité de direction de l'Insee pour donner son aval à la délivrance d'un échantillon, sans inscription au programme des enquêtes de la statistique publique. *« Le Comité de direction a validé le fait que, pour la seule finalité d'obtention d'un échantillon aléatoire de logements, le comité du label puisse émettre, après avoir vérifié la qualité du plan de sondage, de la méthodologie aval et des instances de gouvernance du projet, un simple avis d'examen favorable et non un avis de conformité (délivrance du label d'intérêt général et de qualité statistique, impliquant un visa ministériel et une inscription au JO). De ce fait, les opérations concernées ne seront pas qualifiées d'enquêtes de la statistique publique au sens de la loi de 1951».*

Au final, si l'esprit de l'examen par le Comité du label est depuis sa création de s'assurer de la conformité des méthodes aux règles de l'art, des nuances et des précisions se sont progressivement dessinées dans ses pratiques, conduisant à dresser une typologie de ses avis : avis de conformité, avis d'examen favorable avec visa, avis d'examen simple (sans visa).

UNE PROCÉDURE SPÉCIFIQUE POUR LA LABELLISATION DE STATISTIQUES ADMINISTRATIVES

L'ASP a en charge, entre autres missions, la labellisation de statistiques produites à partir de sources administratives et diffusées par des organismes ayant des missions de service public, donc hors du service statistique public (SSP)[9].

La procédure mise en place pour élaborer ces avis a été définie par l'ASP à la suite d'une mission confiée à l'Inspection générale de l'Insee (Chappert et Puig, 2011), complétée pour sa mise en œuvre opérationnelle par une note d'instruction du Rapporteur du Comité. Pour ces statistiques hors enquêtes du « péri-SSP », la mission proposait de s'appuyer sur le Code de bonnes pratiques de la statistique européenne, en retenant les principes de ce code jugés pertinents pour l'examen de ces données ***(encadré 2)***. Elle proposait également de confier l'instruction des dossiers de labellisation au Comité du label[10] à l'instar de ce qui se fait pour les enquêtes de la statistique publique, ce qui a été acté dans les textes réglementaires en 2013.

Concrètement, l'ASP arrête son programme annuel de statistiques administratives à examiner. Le Comité du label instruit les demandes de labellisation retenues par l'ASP, en recourant, comme dans le cas des enquêtes, à un dossier-type ; il se réunit en commission spécialisée, puis propose un avis assorti de recommandations pour assurer la conformité

9. Voir l'article de Dominique Bureau sur l'ASP, déjà cité, dans ce même numéro.
10. L'ASP peut aussi dans certains cas solliciter une instruction par l'Inspection générale de l'Insee ou d'autres corps d'inspection.

Encadré 2. Critères retenus par l'ASP pour la labellisation de sources administratives

Indépendance, objectivité, impartialité

- Les statistiques sont produites par un service spécialisé, visible dans son organigramme, disposant de moyens humains et financiers appropriés à ses missions statistiques.
- Le responsable de ce service décide en toute indépendance des méthodes d'exploitation ainsi que du contenu et de la date de diffusion des publications.
- Les statistiques sont établies sur une base objective déterminée par des considérations statistiques. Le choix des sources et des techniques se fait en fonction de considérations statistiques.
- Les informations concernant les méthodes et les procédures statistiques suivies sont mises à disposition du public.
- Les dates et heures de parution des statistiques sont annoncées à l'avance.
- Les erreurs découvertes dans les statistiques déjà publiées sont corrigées dans les meilleurs délais et le public en est informé.
- Tous les utilisateurs ont accès aux statistiques dans les mêmes conditions et tout accès privilégié préalable à un utilisateur extérieur est limité, contrôlé et rendu public.
- Les communiqués et déclarations statistiques diffusés dans le cadre de conférences de presse sont objectifs et neutres. Les publications statistiques sont clairement distinguées de la communication de l'organisme sur l'efficacité de son action.

Qualité et pertinence

- Les personnels chargés des exploitations statistiques disposent des compétences nécessaires.
- L'organisme dispose de procédures de gestion et de contrôle de la qualité de sa production statistique transparentes pour les utilisateurs, inspirées des procédures en œuvre dans le SSP.
- Le cadre méthodologique est conforme aux normes, lignes directrices et bonnes pratiques européennes et internationales.
- Les nomenclatures utilisées sont, autant que faire se peut, cohérentes avec celles retenues par le SSP. Les définitions et concepts utilisés à des fins administratives doivent être, dans la mesure du possible, une bonne approximation de ceux qui sont employés en statistique.
- La présentation des résultats ainsi que la périodicité et les délais de publication tiennent compte autant que possible des besoins des utilisateurs.
- Les statistiques sont présentées sous une forme qui facilite une interprétation correcte et des comparaisons utiles.
- Les données collectées, les résultats intermédiaires et les productions statistiques sont évalués et validés.
- Les révisions sont faites selon des procédures normalisées, bien établies et transparentes.
- Les révisions font l'objet d'études et d'analyses qui sont utilisées en interne pour alimenter les processus statistiques.
- Les statistiques sont cohérentes et peuvent être rapprochées pour une durée raisonnable (à cet égard, tout changement dans les règles ou les pratiques administratives susceptibles d'avoir une influence sur les niveaux ou les évolutions doit être porté à la connaissance du public).
- Les métadonnées concernant les méthodes et les procédures suivies ainsi que les résultats sur la qualité statistique des données sont mis à la disposition du public.

des statistiques au Code. La décision de labellisation revient à l'ASP, elle est rendue publique sur son site, ainsi que les conditions ou recommandations qui y sont éventuellement associées et l'avis est publié au Journal officiel.

ET DEMAIN, FAUT-IL QUALIFIER DAVANTAGE DE (SOURCES) STATISTIQUES ?

Avec la démarche de labellisation des enquêtes initiée il y a un peu plus de 25 ans, le Comité du label a développé une méthodologie d'examen permettant d'attester de la qualité des enquêtes de la statistique publique et, plus récemment, des statistiques produites hors du SSP par des organismes ayant des missions de service public. Ces démarches ont constitué des leviers importants pour permettre aux statisticiens d'inscrire leur travail dans des règles respectueuses des bonnes pratiques et d'en assurer la transparence. *In fine,* c'est l'existence même de ce cadre de gouvernance qui constitue un garde-fou, permettant de prévenir autant que possible d'éventuels dysfonctionnements.

Au sein du SSP, les procédures de labellisation continuent de privilégier les enquêtes classiques, alors que d'autres modes de recueils de données ne font pas l'objet du même formalisme. Pour de multiples raisons (dont les moindres coûts de production), on constate en effet un usage croissant de sources administratives ou de données massives pour la production de statistiques par le SSP. La question d'un élargissement de procédures plus formelles de qualification pourrait ainsi se poser au sein du SSP, à l'instar de la labellisation par l'ASP de statistiques administratives produites au sein du « péri-SSP ».

Par ailleurs, avec le développement du numérique, la production de statistiques s'est largement développée hors du SSP : de nombreux acteurs élaborent des bases ou des jeux de données et diffusent des statistiques comme des sous-produits de leur activité principale. Y compris au sein du « péri-SSP », une partie seulement des statistiques produites est d'ailleurs labellisée par l'ASP. De nombreuses statistiques continuent donc d'être reprises par les médias, sans garantie de qualité, ni même de transparence. Un nouveau service de qualification avait été initié au début des années deux-mille-dix avec la démarche « d'étalonnage » par le Cnis de statistiques d'intérêt général produites par des organismes privés. Ce dispositif s'est avéré très coûteux, tant pour les organismes producteurs que pour le Comité lui-même, et n'a pas connu le développement escompté.

De nouvelles réflexions ont conduit à proposer tout récemment un nouveau mode de qualification de statistiques produites hors du SSP : il ne s'agirait plus d'attester de la qualité des statistiques produites, mais seulement de la transparence *a minima* de leur documentation (les « métadonnées »). Cette orientation validée récemment par le Cnis est en cours d'expérimentation. Le principe serait de proposer un standard de documentation, ouvert aux organismes volontaires, les statistiques concernées pouvant ensuite être homologuées par le Cnis. L'objectif est de conduire les utilisateurs à s'interroger sur la portée des données et des statistiques diffusées avant de les utiliser ; il s'agit aussi d'inciter les organismes à documenter leurs produits statistiques pour répondre à ces attentes.

BIBLIOGRAPHIE

ALLAIN, Joël, 1995. Le Comité du label, un an et quelque. In : *Courrier des statistiques.* [en ligne]. Mars 1995. N°73, pp. 63-66. [Consulté le 24 novembre 2020]. Disponible à l'adresse : *https://www.epsilon.insee.fr/jspui/bitstream/1/14343/1/cs73.pdf*.

BONNANS, Dominique, 2019. RMéS, le référentiel de métadonnées statistiques de l'Insee. In : *Courrier des statistiques.* [en ligne]. 27 juin 2019. N°N2, pp. 46-57. [Consulté le 24 novembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4168396/courstat-2-6.pdf*.

CHAPPERT, Alain et PUIG, Jean-Pierre, 2011. *La labellisation de la statistique publique*. [en ligne]. 29 mars 2011. Insee, rapport de l'Inspection générale de l'Insee, n° 80/DG75-B010, Class. 1.6.65. [Consulté le 24 novembre 2020]. Disponible à l'adresse : *https://www.autorite-statistique-publique.fr/rapport-de-linspection-generale-de-linsee-sur-la-labellisation-de-la-statistique-publique/*.

CHRISTINE, Marc, 2016. Assessing and improving quality in official statistics the case of French Label Committee. In : *European Conference on Quality in Official Statistics (Q2016).* [en ligne]. 31 mai – 3 juin 2016. Madrid. [Consulté le 24 novembre 2020]. Disponible à l'adresse : *http://www.ine.es/q2016/docs/q2016Final00118.pdf*.

COMITÉ DU LABEL DE LA STATISTIQUE PUBLIQUE, 2020. *Comité du label de la statistique publique.* [en ligne]. [Consulté le 24 novembre 2020]. Disponible à l'adresse : *https://www.comite-du-label.fr/*.

DESROSIÈRES, Alain, 2003. Les qualités des quantités. In : *Courrier des statistiques.* [en ligne]. Juin 2003. N°105-106, pp. 51-63. [Consulté le 24 novembre 2020]. Disponible à l'adresse : *https://www.epsilon.insee.fr/jspui/bitstream/1/122507/1/cs105-106.pdf*.

ISNARD, Michel, 2018. Qu'entend-on par statistique(s) publique(s) ?. In : *Courrier des statistiques*. [en ligne]. 6 décembre 2018. N°N1. [Consulté le 24 novembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3646978/courstat-1-10.pdf*.

LE GLÉAU, Jean-Pierre *et alii,* 2016. L'obligation de réponse dans la statistique publique. In : *Statistique et société*. [en ligne]. Octobre 2016. Société Française de Statistique (SfdS). Volume 4, n°2, pp. 9-55. [Consulté le 24 novembre 2020]. Disponible à l'adresse : *http://statistique-et-societe.fr/article/download/570/542*.

UNION EUROPÉENNE, 2017. *Code de bonnes pratiques de la statistique européenne.* [en ligne]. 16 novembre 2017. [Consulté le 24 novembre 2020]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/documents/4031688/9332274/KS-02-18-142-FR-N.pdf/130905e7-45a7-4475-b37c-8f699b5e33e1*.

VORMS, Bernard, JACQUOT, Alain et LHÉRITIER, Jean-Louis, 2010. *L'information statistique sur le logement et la construction.* [en ligne]. 16 mars 2010. Cnis, rapport d'un groupe de travail du Cnis n°121. [Consulté le 24 novembre 2020]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2010_121_logement_construction.pdf*.

RÉFÉRENCES JURIDIQUES

Arrêté du 10 janvier 1994 portant création au sein du Conseil national de l'information statistique d'un comité du label des enquêtes statistiques des services publics et des autres services producteurs d'informations statistiques. In : *site de Légifrance.* [en ligne]. Modifié le 24 décembre 1997. [Consulté le 24 novembre 2020]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000543608/1997-12-23/*.

Arrêté du 2 mai 2013 relatif aux modalités d'organisation du comité du label de la statistique publique. In : *site de Légifrance.* [en ligne]. [Consulté le 24 novembre 2020]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000027412247/*.

Décret n° 2009-318 du 20 mars 2009 relatif au Conseil national de l'information statistique, au comité du secret statistique et au comité du label de la statistique publique. In : *site de Légifrance.* [en ligne]. Modifié le 1er février 2019.[Consulté le 24 novembre 2020]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000020428769*.

Loi n° 2008-776 du 4 août 2008 de modernisation de l'économie. In : *site de Légifrance.* [en ligne]. Modifiée le 5 juillet 2019. [Consulté le 24 novembre 2020]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000019283050/2020-11-20/*.

Loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques. In : *site de Légifrance.* [en ligne]. Modifiée le 25 mars 2019. [Consulté le 24 novembre 2020]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573/2020-11-20/*.

LES DONNÉES CARROYÉES, DES OUTILS ET MÉTHODES INNOVANTS

POUR PERCEVOIR LA RÉALITÉ DES TERRITOIRES

Valérie Darriau[*]

Les données carroyées sont des données diffusées sur une maille originale, ne correspondant à aucun découpage administratif ou historique connu : celle de carrés, dont les côtés peuvent aller de 200 mètres jusqu'à plusieurs kilomètres. Dans les zones urbaines, quand les découpages communaux sont trop imprécis pour analyser les phénomènes démographiques ou socio-économiques, l'assemblage des carreaux permet de fournir des informations précieuses.

Pour produire ce type de données, l'Insee doit relever plusieurs défis : géolocaliser les informations pour les rattacher à des carreaux, développer une méthode garantissant la protection de la vie privée et le respect du secret statistique, mettre à disposition ces données sous une forme utilisable par des experts, mais également par des amateurs, curieux de mieux connaître leur territoire et d'en avoir un aperçu rapide et éclairant.

Avec quelques exemples d'utilisation pour le déploiement de politiques publiques, l'article illustre les techniques mises en œuvre pour permettre la diffusion des données issues de sources socio-fiscales en 2019. À l'image d'une mosaïque, un carreau pris individuellement n'a pas de sens : c'est bien la proximité avec ses voisins qui va permettre à la réalité de prendre forme et au territoire de révéler la richesse et la complexité des phénomènes qui le traversent.

Grid data are data disseminated on an original lattice: that is not corresponding to any known administrative or historical division, but squares, whose sides can range from 200 metres to several kilometres. In urban areas, when municipal boundaries are too imprecise to analyse demographic or socio-economic phenomena, the assembly of grid can provide valuable information.

In order to produce this type of data, INSEE has to meet several challenges: geolocate the information to attach it to grids, develop a method that guarantees the protection of privacy and respect for statistical secrecy, and make this data available in a form that can be used by experts, but also by amateurs, who are curious to get to know their territory better and to have a quick and enlightening overview of it.

With a few examples of use for the deployment of public policies, the article illustrates the techniques that have been implemented to enable the dissemination of data from socio-fiscal sources in 2019. Like in a mosaic, a single "tile" does not make sense: it is indeed the proximity with its neighbours that will allow reality to take shape and the territory to reveal the richness and complexity of the phenomena that run through it.

* À la date de la rédaction de l'article, cheffe de la division Statistiques et analyses urbaines, Insee, *valerie.darriau@insee.fr*

En 2013, l'Insee a diffusé un premier jeu d'indicateurs sur une nouvelle maille géographique : le carreau. Cette diffusion a rencontré un succès important auprès des agences d'urbanisme et spécialistes des analyses urbaines. L'attente de la mise à jour et de l'enrichissement de ces données était forte. En 2019, l'Insee a répondu à cette demande, et a souhaité faciliter l'accès à ces données et leur utilisation. Les enjeux qui se cachent derrière leur production sont plus complexes qu'il n'y paraît. Avant toute chose, les données carroyées nécessitent une méthodologie permettant de garantir confidentialité et qualité des données. Le département des Méthodes statistiques de l'Insee a engagé des travaux innovants pour répondre à cette exigence (Branchu, Costemalle et Fontaine, 2018 ; Costemalle, 2018). L'écoute des utilisateurs a également eu un rôle important pour mettre à disposition des experts de l'analyse urbaine comme du grand public les données sous la forme la plus adaptée à leurs besoins.

L'ÎLOT ET L'IRIS, PREMIERS DÉCOUPAGES STATISTIQUES DU TERRITOIRE COMMUNAL

La connaissance d'un territoire nécessite de mobiliser des données statistiques à une échelle géographique fine. Très souvent, c'est le niveau communal qui est utilisé comme brique de base pour constituer des zonages d'études et répondre à des problématiques spécifiques : bassins de vie, zones d'emploi, unité urbaine, et plus récemment aires d'attraction des villes (de Bellefon, Eusebio, Forest, Pégaz-Blanc et Warnod, 2020). Les communes forment en effet une partition du territoire national, et disposent d'une offre de données riche, en particulier grâce au recensement de la population.

Cependant, dans les villes et agglomérations, les périmètres de l'intervention publique ou des projets territoriaux ne correspondent que rarement aux seules frontières des communes et nécessitent de disposer d'informations à des mailles encore plus fines.

Jusqu'en 1999, l'Insee a diffusé des données statistiques issues de chaque recensement au niveau de l'îlot[1], périmètre qui correspondait à un pâté de maison. En parallèle, dans les années quatre-vingt-dix, l'Insee a construit avec les plus grandes communes un maillage de diffusion appelé IRIS[2], utilisé notamment pour les variables « sensibles » du recensement : ces périmètres regroupaient des îlots contigus d'une même commune, en dessinant des sortes de « quartiers ».

Utiles pour l'analyse urbaine fine, le découpage du territoire communal en îlots variait cependant entre chaque recensement, rendant malaisée, voire impossible, l'analyse des évolutions des phénomènes urbains. Depuis le recensement rénové, en 2004, la collecte du recensement n'est plus exhaustive dans les grandes communes mais se fait par sondage. Pour garantir la robustesse des données diffusées, la diffusion infracommunale a abandonné la maille de l'îlot pour adopter la seule échelle des IRIS ; avec les évolutions du territoire urbain, les contours des IRIS ont progressivement évolué, pour ne plus être systématiquement emboîtés avec la dernière maille à l'îlot, celle du recensement 1999.

1. Voir la définition de l'îlot sur le site de l'Insee (*https://www.insee.fr/fr/metadonnees/definition/c1656*).
2. Des IRIS (Îlots regroupés pour l'information statistique), de population moyenne égale à 2 000 habitants, ont alors été définis pour toutes les communes de plus de 5 000 habitants. Pour plus de précision, voir le site de l'Insee (https://www.insee.fr/fr/metadonnees/definition/c1523).

LE CARREAU, UNE MAILLE NEUTRE, «SIMPLE ET PRATIQUE» POUR L'ANALYSE URBAINE

Pour chaque grande thématique, qu'il s'agisse de l'habitat, des transports, des équipements de proximité ou de la santé, l'espace urbain s'analyse selon des découpages spécifiques. Pour identifier, par exemple, la population résidant à proximité d'une gare, le long d'une infrastructure ou encore exposée au bruit, les périmètres d'analyse se doivent d'être précis. Leur approximation par les mailles îlot ou IRIS ne satisfont pas complètement les acteurs locaux. Leurs contours ne sont pas stables dans le temps, et leur géométrie est variable : petits en centre-ville, ils sont beaucoup plus étendus en périphérie. De plus, ces contours irréguliers entraînent l'apparition d'un phénomène géographique connu sous l'acronyme MAUP[3] : les formes irrégulières et les limites des maillages administratifs qui ne reflètent pas nécessairement la réalité des distributions spatiales étudiées sont un obstacle à la comparabilité des unités spatiales inégalement subdivisées (Loonis et de Bellefon, 2018).

Dans les années quatre-vingt, une technique nouvelle se développe alors, celle du carroyage (Delahaye, 1987). Son principe ? Découper le territoire en petits carrés de taille identique sur lesquels on va créer de l'information qu'il n'y aura alors plus qu'à agréger sur le territoire d'intérêt. Son usage apparaît « *simple et pratique* » (Certu et Cete Normandie Centre, 2011), il permet des comparaisons spatiales et temporelles plus aisées.

> *« Plus le découpage est fin, plus les carreaux sont petits et nombreux, plus la taille des bases de données est volumineuse et les capacités de traitement informatique sont alors inadaptées. »*

La difficulté réside toutefois dans la disponibilité des données : comment disposer, sur ces carreaux, de données utiles pour l'analyse, alors qu'elles sont collectées en général au niveau de la commune ou d'îlots de taille différente, de géométrie variable ? Des méthodologies se développent alors pour désagréger ou répartir l'information disponible à l'échelle de quartiers sur des carreaux (Lajoie, 1992). Mais plus le découpage est fin, plus les carreaux sont petits et nombreux, plus la taille des bases de données est volumineuse et les capacités de traitement informatique sont alors inadaptées. L'évolution des technologies informatiques, le développement de la géolocalisation infracommunale et des systèmes d'information géographique ont permis de pallier ces difficultés et de démocratiser cette représentation des données.

LE CARREAU, BRIQUE DE BASE STABLE POUR CONSTRUIRE UN ZONAGE D'INTÉRÊT

Fruit du découpage d'un territoire selon une grille régulière, chaque carreau pris individuellement n'a pas de « sens » géographique, il ne reflète aucune réalité territoriale connue. En revanche, il permet de reconstituer une zone d'intérêt en étant agrégé avec ses voisins.

3. *Modifiable Areal Unit Problem.*

Cette méthode a ainsi été utilisée par l'Observatoire des quartiers de gare mis en place par l'Atelier parisien d'urbanisme (Apur), pour caractériser les quartiers des futures gares du cœur de l'agglomération parisienne. Pour analyser le quartier des Ardoines, matérialisé par un disque de 800 mètres autour de la gare de Vitry-sur-Seine, l'Apur a utilisé deux approches ***(figure 1)***. La première consiste à mobiliser les données carroyées en sélectionnant les carreaux qui intersectent (même de façon minime) le périmètre. Une enveloppe de 63 carreaux est ainsi créée : sa surface totale (2,5 km²) est proche de celle du disque (2,0 km²). La seconde approche de l'Apur est de faire appel aux IRIS qui intersectent le disque. Ils sont de périmètres très variables et forment une surface totale de 3,7 km², soit près du double du quartier initialement analysé. Le carreau permet ici de reconstituer une statistique plus proche de la réalité du quartier que ne le peut l'IRIS.

L'autre intérêt de former la zone d'observation à partir des carreaux est que cela garantit une analyse temporelle sur un périmètre constant, utile pour des analyses sur longue période.

DES CARREAUX, MAIS LESQUELS ?

Au début des années 80, en Grande-Bretagne, un atlas présente les résultats du recensement de 1971 à l'échelle de carreaux de 1 km² et 10 km² (OPCS, 1980). Dans d'autres pays, les données carroyées se développent, dans différentes disciplines (médecine, géologie, botanique, biologie…). Mais les tailles des carreaux et l'emplacement de la grille utilisée sont variables et propres aux territoires analysés.

Au travers de la directive européenne **INSPIRE** (Directive 2007/2/CE), l'Union européenne a souhaité établir *« une infrastructure de données géographiques pour assurer l'interopérabilité entre bases de données et faciliter la diffusion, la disponibilité, l'utilisation et la réutilisation de l'information géographique en Europe »*. S'agissant des carreaux, elle vise notamment à promouvoir la création d'une *« grille multi-résolution harmonisée avec un point d'origine commun et une localisation ainsi qu'une taille des cellules harmonisées »* (Cnig, 2020). Ainsi, la « France carroyée » peut-elle se juxtaposer à l'Allemagne ou à l'Italie carroyée, selon un schéma standard et compatible.

Pour trouver un carreau de façon univoque, il faut d'abord connaître la taille de la grille dans laquelle on se situe. Le carreau aura-t-il une longueur de 200 mètres de côté ou de 4 km ? Il faut que son identifiant précise cette information (European Commission, 2010). Une fois cette résolution connue, INSPIRE a imposé par convention d'identifier un carreau par son coin inférieur gauche. En effet, avec ce seul point et la résolution de la taille du carreau (200 m), on peut tout de suite tracer le carreau correspondant, en partant du coin et en allant vers l'Est sur 200 m et vers le Nord sur 200 m. Ces deux éléments suffisent donc théoriquement à trouver le carreau, à une subtilité près : les coordonnées géographiques dépendent en effet de la projection[4] utilisée. L'identifiant INSPIRE doit donc aussi le spécifier ***(figure 2)***.

Une fois les carreaux identifiés, l'étape suivante consiste à les relier aux informations statistiques.

4. En cartographie, un système de projection (ou de coordonnées) est un référentiel dans lequel on peut représenter des éléments dans l'espace. Ce système permet de se situer sur l'ensemble du globe terrestre grâce à un couple de coordonnées géographiques.

CARREAUX + DONNÉES = DONNÉES CARROYÉES ?

L'équation semble facile à réaliser, mais sa résolution n'est pas si simple. Où résident les populations en grande précarité ? Les personnes âgées ? Où se trouvent les logements anciens ? La réponse à ces questions est présente dans des fichiers fiscaux, administratifs, de gestion… Il faut traduire une information individuelle et fine en une donnée agrégée sur l'ensemble du carreau. Pour cela, les fichiers doivent contenir une indication géographique permettant de rattacher précisément l'information à un carreau. Certains disposent déjà de coordonnées géographiques précises : c'est le cas des fichiers d'origine fiscale (taxe d'habitation par exemple) qui contiennent, outre des informations statistiques (population, revenu, etc.), un identifiant de la parcelle cadastrale sur laquelle se situe le logement ou le foyer fiscal concerné.

Figure 1. Les carreaux permettent d'approcher un périmètre d'intérêt plus finement que les IRIS

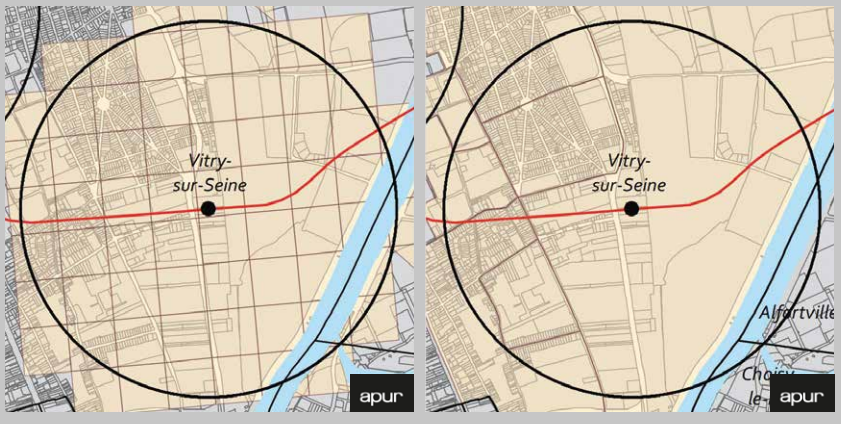
L'Observatoire des quartiers de gare du Grand Paris a été mis en place par l'Atelier parisien d'urbanisme (Apur). Dans l'exemple choisi, le quartier des Ardoines est ici matérialisé par un disque de 800 mètres autour de la gare de Vitry-sur-Seine.

Pour comparer l'efficacité des découpages au carreau ou à l'IRIS, l'Apur a utilisé deux approches : le carreau (à gauche) permet ici de reconstituer une statistique plus proche de la réalité du quartier que ne le peut l'IRIS (à droite).

63 carreaux, de couleur beige, intersectent le disque. L'enveloppe ainsi créée a une surface totale (2,5 km²) très proche de celle du disque (2,0 km²).

Les IRIS qui intersectent le disque sont de périmètres très variables et forment une surface totale de 3,7 km², soit près du double du quartier initialement analysé.

Exemple du quartier des Ardoines :



Découpage du quartier au carreau de 200 m x 200 m.

Découpage du quartier à l'IRIS

Cartes tirées de (Apur, 2014)

Comme dans l'exemple de la ***figure 3***, chaque parcelle possède par ailleurs une étiquette, matérialisée par un point dont les coordonnées sont connues. Ce point est en général situé dans la parcelle : il suffit donc de rattacher ce point au carreau qui le recouvre pour positionner les informations de la parcelle dans le carreau.

Simple en apparence, l'opération peut se révéler quelquefois complexe. Dans de très rares cas, l'étiquette d'une parcelle peut être positionnée à l'extérieur de la parcelle, voire de la commune. Il faut alors la repérer et, si possible, corriger sa position. Selon le découpage utilisé, il peut aussi arriver que l'information soit localisée dans un carreau qui ne recouvre qu'une toute petite partie de la parcelle. Ainsi, dans l'exemple de la ***figure 3***, le carreau n°1 s'étend sur les parcelles n°0048 et 0383, ainsi une partie de la parcelle n°049 qui comprend le bâti supposé habité ; mais cette dernière parcelle est rattachée au carreau voisin, le n°4. Les informations statistiques du carreau n°1 ne correspondront *in fine* qu'à une seule habitation (située sur la parcelle n°048), alors qu'il en recouvre deux (celle de la parcelle n°48 et celle de la parcelle n°49).

LOCALISER L'INFORMATION DANS LE BON CARREAU : LE DÉFI DE L'ADRESSE POSTALE

Le plus souvent, les fichiers statistiques ne contiennent comme information géographique qu'une adresse postale : celle des allocataires des prestations familiales, des logements sociaux, etc. Il faut alors la situer précisément sur une carte, pour pouvoir rattacher les données correspondantes à un carreau.

Pour ce faire, il faut géolocaliser l'adresse, c'est-à-dire reconnaître la chaîne de caractères de l'adresse postale dans un « référentiel », sorte de répertoire qui contient les numéros et libellés des rues de toutes les communes de France, mais aussi leur emplacement précis. Une fois la chaîne de caractères retrouvée, on rattache alors à l'information statistique du fichier les coordonnées géographiques présentes dans le référentiel.

En milieu urbain, les adresses sont très souvent normalisées : elles comportent un numéro, un type de voie (avenue, rue…), un libellé de voie et un code postal ou un code commune. Dans ces cas, l'adresse postale est univoque. Les difficultés tiennent alors à l'identification correcte des chaînes de caractères : par exemple, une abréviation contenue dans le libellé *« 245 rue du Dr Fiolle (à Marseille) »* doit pouvoir être rattachée au libellé du référentiel *« 245 rue du Docteur Fiolle ».*

> *« En milieu urbain, les adresses sont très souvent normalisées. En milieu rural, les difficultés sont souvent d'une autre nature. »*

En milieu rural, les difficultés sont souvent d'une autre nature, car les adresses ne sont pas toutes normalisées : c'est très souvent le cas des lieux-dits. Retrouver de façon certaine la localisation d'un ménage ***(figure 4)*** qui indique comme adresse *« Bussac Bas, hameau de Siaugues-Sainte-Marie »,* est quasi impossible car plusieurs maisons utilisent ce libellé d'adresses. Il faudra donc choisir de façon arbitraire à quelle adresse rattacher l'information statistique correspondante.

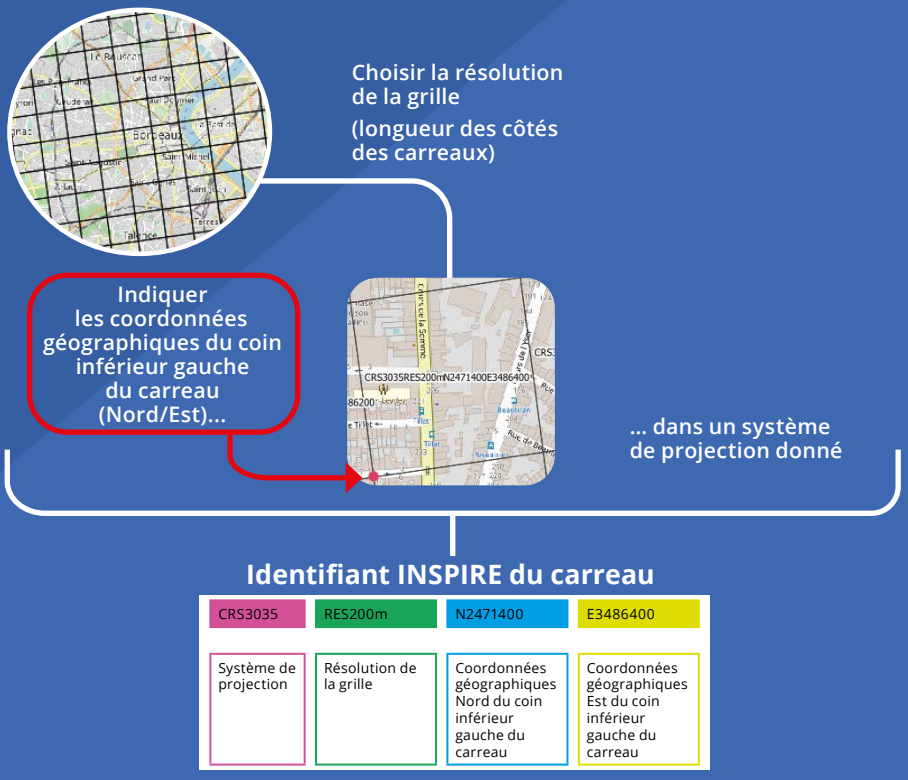
Quel que soit le procédé employé, utilisation de l'étiquette de parcelle cadastrale ou géolocalisation d'adresses postales, les informations statistiques sont *in fine* localisées précisément sur un plan et rattachées au carreau correspondant. On obtient alors des bases de données, appelées **données carroyées :** ce sont ces bases qui vont servir ensuite aux acteurs publics et privés à éclairer des problématiques spécifiques.

LES DONNÉES CARROYÉES, UTILES POUR GUIDER LA DÉCISION PUBLIQUE

De la Commission européenne aux collectivités locales françaises, chacun plaide pour disposer d'informations précises sur l'urbain.

Au niveau européen, selon Lewis Dijkstra[5], les données carroyées *« présentent de nombreux avantages. Parce qu'elles fournissent des données à haute résolution spatiale dans une forme et une taille standardisées, ces données peuvent être combinées de manière transparente avec les données des pays voisins. La Commission européenne s'appuie fortement sur les données carroyées pour l'analyse de l'accès aux services, comme les transports, l'éducation et les soins de santé, y compris aux services implantés de l'autre côté de la frontière nationale. En outre, les données carroyées jouent un rôle essentiel dans l'évaluation de l'exposition à la pollution et aux risques naturels et peuvent aider à orienter les services d'urgence ».*

Figure 2. Comment INSPIRE identifie les carreaux



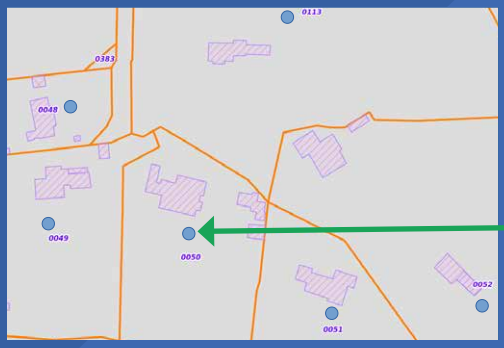
NB : Afin de couvrir l'ensemble du territoire européen, la projection utilisée conduit à visualiser des carreaux légèrement « penchés » sur le territoire français.
Voir également (European Commission, 2010).

5. Lewis Dijkstra est le chef adjoint de l'unité Développement politique et analyse économique à la Direction générale de la politique régionale et urbaine de la Commission européenne.

Au niveau local, les premiers utilisateurs de données carroyées sont les agences d'urbanisme : *« Grâce à la finesse du maillage du carroyage, l'Agence d'urbanisme de Bordeaux Aquitaine a pu produire des analyses territorialisées riches et utiles. »,* indique Caroline De Vellis[6]. *« Dans le Schéma directeur opérationnel des déplacements métropolitains, des éléments clés ont été apportés sur la population disposant d'un service de transport en commun à proximité. Nous avons aussi pu qualifier la densité de population sur l'étude du RER métropolitain, ou encore analyser et comparer des quartiers jusque-là mal identifiés par le découpage en IRIS ».*

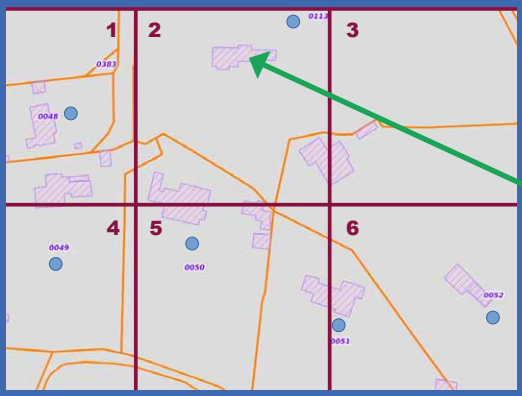
Figure 3. Carroyage de données fiscales associées à une parcelle cadastrale

Fichier statistique nativement localisé



Parcelle	Population	Revenus	Individus de moins de 5 ans
0383	0	0	0
0048	2	40 000	0
0049	6	30 000	2
0050	4	60 000	0
0051	5	50 000	2
0052	4	40 000	1
0113	1	35 000	0

Passage au carreau : les points d'étiquettes sont rattachés à chaque carreau



Carreau	Parcelles	Population	Revenus	Individus de moins de 5 ans
1	0383, 0048	2	40 000	0
2	0113	1	35 000	0
3	-	-	-	-
4	0049	6	30 000	2
5	0050	4	60 000	0
6	0051, 0052	9	90 000	3

6. Caroline de Vellis est statisticienne à l'Agence d'urbanisme de Bordeaux Métropole, animatrice du club Observation de la Fédération nationale des agences d'urbanisme (FNAU).

D'autres acteurs sont friands de ces informations : bureaux d'études, chercheurs, étudiants, collectivités locales moins outillées pour manipuler les données, tous réclament en outre des données facilement accessibles, des supports cartographiques et des outils leur permettant de les manipuler facilement.

CONCILIER RICHESSE DE L'INFORMATION…

Les données diffusées au carreau par l'Insee sont pour l'instant limitées à la source Filosofi[7]. Mobilisant des informations d'origine fiscale et sociale, cette source construite à des fins statistiques permet de fournir des indicateurs de niveau de vie, d'inégalité et de pauvreté mais aussi des données socio-démographiques, à un niveau local fin, répondant à une partie des besoins des utilisateurs. Ces données permettent ainsi d'éclairer les thématiques démographiques (petite enfance, personnes âgées, etc.), sociales (pauvreté, familles monoparentales), scolaires (fréquentation des écoles, collèges, etc.), environnementales (ancienneté des logements), urbaines (logement social, accession à la propriété), etc.

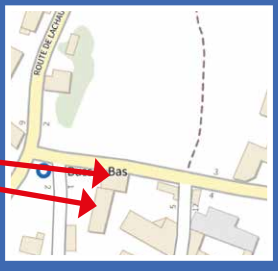
Figure 4. La géolocalisation des adresses postales est plus complexe en espace rural

N° Logement	adresse	Revenus	Individus de moins de 5 ans	Indi à 10
1	27 boulevard des fleurs, 33000 Bordeaux	0111	11	
2	350 avenue Léon Blum, 14302 Caen	0216	[illegible]	
3	40 rue Denuzière, 69002 Lyon	0333	33	
4	Bussac Bas, 43300 Siaugues-Saintes-Marie	0444	44	

En zone urbaine, les adresses sont souvent normalisées et les échos sont univoques

N° Logement	adresse	Revenus	Indivic de mo de 5 a
1	27 boulevard des fleurs, 33000 Bordeaux	0111	11
2	350 avenue Léon Blum, 14302 Caen	0121	22
3	40 rue Denuzière, 69002 Lyon	0333	33
4	Bussac Bas, 43300 Siaugues-Saintes-Marie	[illegible]	44

?



En zone rurale, les échos sur les lieux-dits sont nombreux, et la géolocalisation moins efficace et précise

7. Filosofi désigne le dispositif sur les revenus localisés sociaux et fiscaux de l'Insee.

Grâce à cette source, il est possible de disposer sur un carreau de nombreuses informations :

- informations sur les individus (nombre, tranches d'âge, etc.) ;
- informations sur les ménages (nombre, taille, niveaux de vie, statut de propriétaire, familles monoparentales, etc.) ;
- caractéristiques des logements (logements collectifs, logements sociaux, maisons, dates de construction).

La tentation est alors forte de vouloir croiser ces informations, pour disposer, par exemple, du nombre de ménages pauvres en logement social. Mais la finesse de la maille de diffusion (carreau de 200 m) impose des précautions particulières pour protéger la vie privée et respecter le secret statistique et le secret fiscal.

… ET GESTION DU SECRET STATISTIQUE

Le secret statistique concerne la protection des individus de toute diffusion de données individuelles et de toute ré-identification à partir des données statistiques.

Le secret fiscal régit de son côté l'utilisation des données issues de la source fiscale, donc de la source Filosofi utilisée par l'Insee. Il impose que les informations statistiques soient diffusées uniquement sur des agrégats d'au moins 11 ménages fiscaux.

Pour respecter ces dispositions, le département des Méthodes statistiques de l'Insee a élaboré une méthodologie de carroyage spécifique et originale (Branchu, Costemalle et Fontaine, 2018) qui aboutit à une diffusion des données carroyées sur deux types de grilles différents (Insee, 2019).

En effet, les grilles de carreaux évoquées jusqu'ici étaient implicitement des grilles « régulières », c'est-à-dire dont les tailles de carreaux étaient identiques partout. Mais dans l'optique de la gestion de la confidentialité, la méthode employée amène à revoir cette hypothèse. Car si, en milieu urbain, le seuil de 11 ménages est très souvent respecté sur un carreau de 200 m, en revanche, dans les zones moins denses, il est plus difficile à atteindre : 79 % des carreaux de France métropolitaine, de Martinique et de La Réunion comprennent ainsi moins de 11 ménages. Il faut parfois atteindre une taille de carreau de 32 km pour couvrir suffisamment de zones habitées et que l'information diffusée à cette échelle garantisse la confidentialité.

Une première méthode dite de « niveau naturel » va donc adapter la taille du carreau au nombre d'habitants qu'il contient. Le second type de grille est celui, plus intuitif, des grilles régulières : pour gérer la confidentialité sur ce type de grille, il faut alors accepter que les données de certains carreaux soient modifiées.

LA GRILLE AUX CARREAUX DE TAILLE DIFFÉRENTE : LE NIVEAU «NATUREL»

La grille de niveau naturel correspond à un partitionnement du territoire en carreaux de différentes tailles (de 200 m jusqu'à 32 km) permettant de diffuser toutes les informations, tout en respectant le secret fiscal.

Concrètement, on commence par couvrir le territoire avec des carreaux de 32 km, taille nécessaire pour être certain que dans chacun de ces carreaux, il y a au moins 11 ménages. Puis on les divise en 4, pour former des carreaux de 16 km dans lesquels on décompte le nombre de ménages présents. Si jamais l'un d'eux abrite moins de 11 ménages, alors la grille ne sera pas découpée à ce niveau. On poursuit et les divisions s'arrêtent :

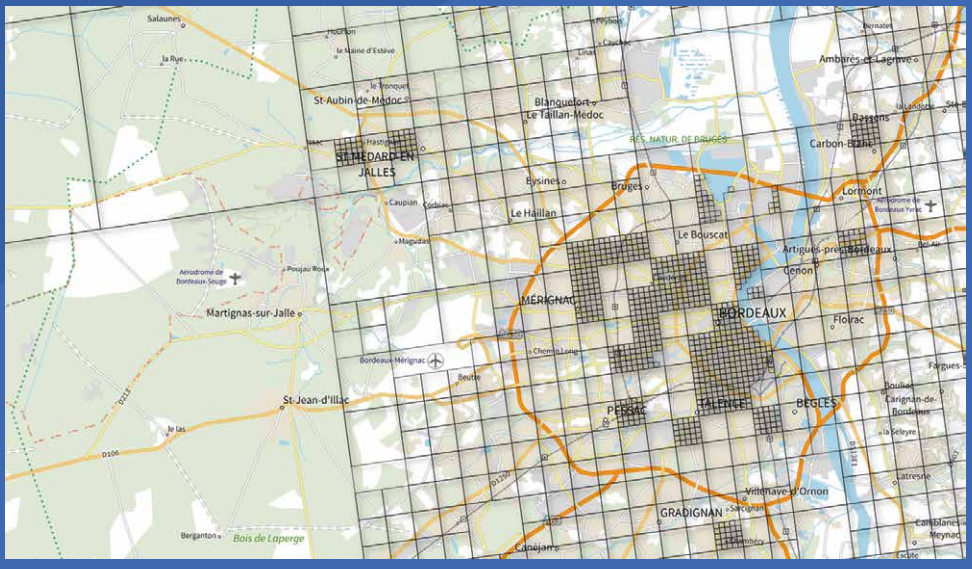
- soit lorsque les carreaux obtenus sont de taille 200 m ;
- soit lorsque la prochaine division entraînerait qu'un ou plusieurs carreaux ne respectent pas le seuil de confidentialité fixé à 11 ménages.

Dans les territoires peu denses, la division s'arrête tôt, sur des carreaux de taille élevée, comme on le voit dans l'exemple de la ***figure 5***, à l'ouest de l'agglomération bordelaise. Dans les territoires très denses, comme le centre-ville, les données seront disponibles à 200 m.

Ce premier niveau de diffusion permet de garantir l'exactitude de toutes les données diffusées sur chaque carreau. Néanmoins, il ne se prête pas vraiment à une représentation cartographique des données : les carreaux peu denses et peu peuplés, ont une superficie très importante qui captera l'attention visuelle alors que les carreaux denses du centre-ville seront peu visibles, renforçant l'effet du MAUP cité plus haut (Floch, 2012).

En outre, il est dépendant de la source et changera donc si on diffuse d'autres sources statistiques – ou un autre millésime de cette source – au carreau. On ne pourra donc pas superposer les grilles de niveau naturel de deux sources différentes.

Figure 5. Exemple de découpage en carreau de niveau «naturel» dans l'agglomération bordelaise



LA GRILLE DE CARREAUX DE 200 MÈTRES… OU LA «RÈGLE DU 80/20»

Plus familier, le deuxième type de grille consiste à proposer un découpage régulier, de taille de carreau fixe.

Cette grille offre plusieurs avantages. Tout d'abord, elle permet de disposer d'un découpage utilisable pour n'importe quelle source. Elle permet également de récupérer de l'information disponible à un niveau géographique plus fin que le niveau naturel ne l'autorise. En effet, si le découpage du niveau naturel garantit la diffusion de données exactes, il n'optimise pas l'information diffusée.

Prenons l'exemple fictif de la ***figure 6***. Le carreau de 1 km comprend 555 ménages. Mais lors du découpage à 200 mètres, 14 carreaux ayant moins de 11 ménages sont identifiés (en orangé) : le niveau naturel est donc dans ce cas le carreau de 1 km.

Cependant, on voit qu'au niveau du découpage à 200 m, 11 carreaux sont supérieurs au seuil et rassemblent 450 ménages, soit 81 % du nombre total. L'information pourrait être diffusée sur ces carreaux-là sans trahir le secret, or le niveau naturel ne le permet pas.

En contrepartie, il faut traiter l'information présente dans les carreaux de moins de 11 ménages. La première option envisagée pourrait être de les «blanchir», c'est-à-dire de ne pas diffuser les valeurs de ces carreaux. Toutefois, cela impliquerait que les valeurs du carreau de 1 km soient différentes de la somme des valeurs des carreaux de 200 m le composant. La seconde option est donc de récupérer l'information des carreaux non diffusables et de la répartir « aléatoirement » entre eux au sein du carreau de 1 km. Ce procédé garantit ainsi un gain d'information, la cohérence des totaux entre les niveaux de diffusion mais se traduit par la présence de données modifiées sur les carreaux non diffusables. Il est alors impératif que l'utilisateur soit averti de la méthode et qu'il puisse distinguer les valeurs réelles des valeurs imputées. Dans le fichier des données carroyées de la France métropolitaine, 80 % des carreaux de 200 m font l'objet d'une imputation, mais ils ne représentent que 20 % de la population.

LES VARIABLES SENSIBLES : PAUVRETÉ ET NIVEAUX DE VIE

La méthodologie élaborée permet de s'assurer qu'aucune information portant sur moins de 11 ménages n'est diffusée. Néanmoins, s'agissant des informations sur la pauvreté et les revenus, l'Insee a souhaité appliquer des précautions supplémentaires :

- pour les carreaux ayant plus de 11 ménages mais dont plus de 80 % des ménages sont pauvres : le chiffre du nombre de pauvres a été ramené à la valeur de 80 % ;
- pour la distribution des niveaux de vie, les valeurs extrêmes ont également fait l'objet d'un traitement particulier, appelé *winsorisation,* qui permet d'éviter la sensibilité aux valeurs extrêmes de la distribution. En pratique, après avoir calculé les niveaux de vie de chaque individu, on regarde la distribution de ces niveaux de vie pour un département donné :
  - si le niveau de vie d'un individu est supérieur au 95e centile de la distribution départementale, son niveau de vie est rabaissé à ce seuil [par exemple, dans l'Ain, si un individu a un niveau de vie de 60 000 € annuel, on lui affecte la valeur 54 680 €] ;
  - inversement, si son niveau de vie est inférieur au 5e centile de la distribution départementale, son niveau de vie est ramené à ce seuil [toujours dans l'Ain, si un individu a un niveau de vie de 8 000 € annuel, on lui affecte la valeur 9 010 €] ;
  - si son niveau de vie se situe entre ces deux seuils, aucun traitement n'est effectué.

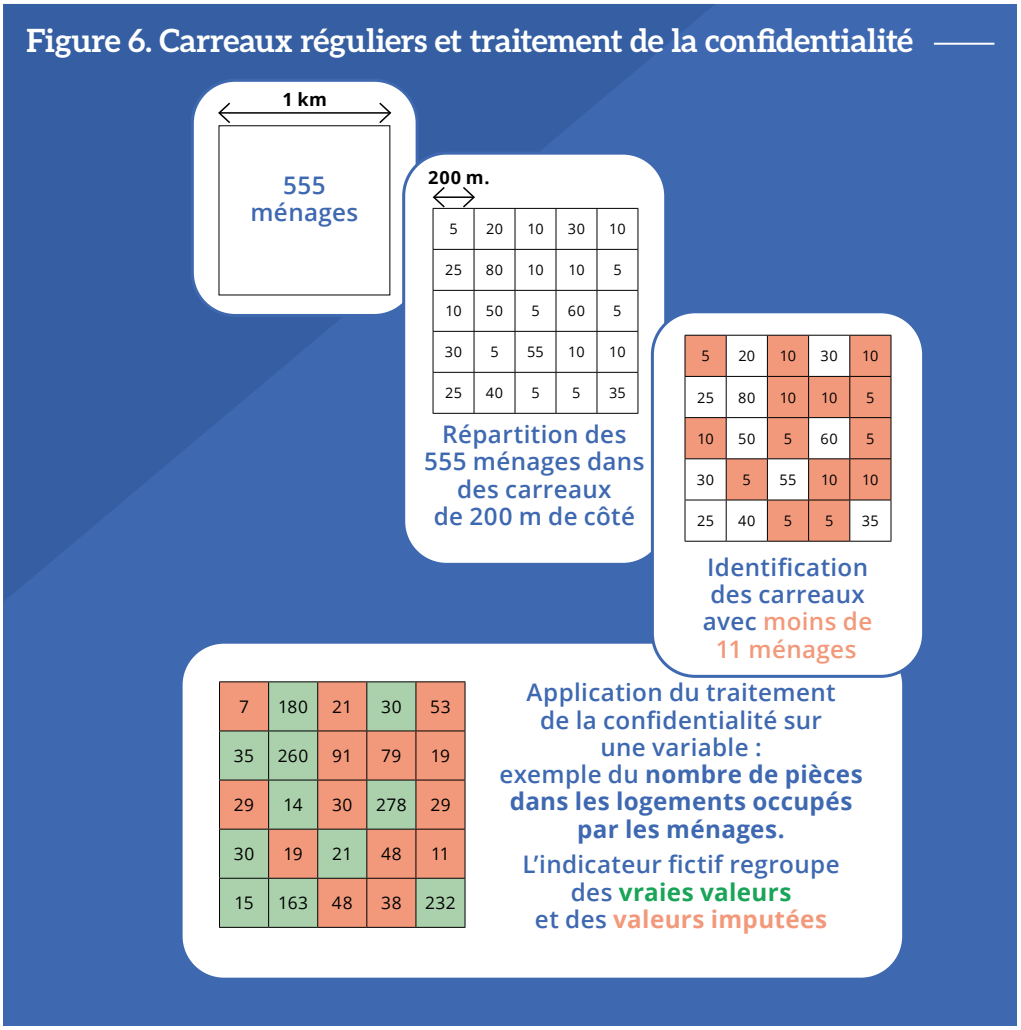
Ce traitement permet de protéger les informations individuelles, tout en préservant l'information utile à l'analyse territoriale.

L'ensemble des traitements méthodologiques a fait l'objet d'une déclaration au délégué à la protection des données dont l'Insee relève, et les modalités de protection des données personnelles sont accessibles sur le site de l'Insee (Insee, 2020a).

POUR L'UTILISATEUR AVERTI, DES BASES DE DONNÉES POUR EXPRIMER SA CRÉATIVITÉ

Une fois les données prêtes à l'emploi, positionnées dans chaque carreau et « secrétisées », il ne reste qu'à… les utiliser. Certains utilisateurs sont experts en traitement de données et logiciels de cartographie : ils souhaitent disposer des données brutes qu'ils pourront ensuite manipuler à leur guise, pour les représenter de la façon la plus appropriée à leurs besoins. Ils apprécient la souplesse offerte par les données, qui forment autant de briques que l'on peut assembler pour créer des représentations spatiales originales.

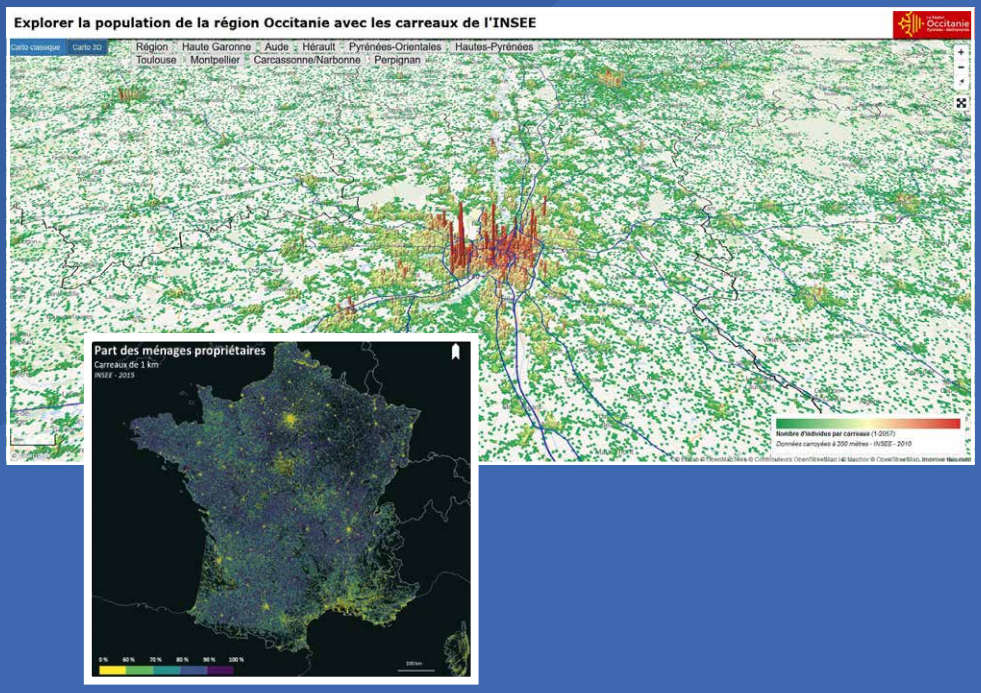
Figure 6. Carreaux réguliers et traitement de la confidentialité



Pour ces spécialistes, des formats de diffusion adaptés ont été utilisés, comme le format *shapefile,* très répandu dans l'analyse cartographique mais « propriétaire » ou le format *geopackage,* plus volumineux mais libre.

Boris Mericskay[8], auteur des représentations cartographiques de la ***figure 7***, souligne que *« ces données très originales permettent de donner à voir et d'appréhender à différentes échelles (du pays au quartier en passant par commune) les dynamiques territoriales de peuplement, de dynamisme socio-économique comme le revenu médian, les tranches d'âge ou les dates de construction des logements. La modélisation proposée, à savoir un maillage régulier du territoire permet aussi d'explorer de nouvelles formes de représentations et de géovisualisation de données géographiques issues de la statistique publique ».* En revanche, ces bases de données restent lourdes à télécharger et complexes à manipuler. *« Le seul bémol que je peux souligner réside dans la mise à disposition de ces données par l'Insee sous un fichier unique, trop lourd pour la manipulation de non-spécialistes »* confirme Boris Mericskay.

Figure 7. Quelques utilisations des données carroyées par les internautes



Cartes réalisées par B. Mericskay

https://www.sites.univ-rennes2.fr/mastersigat/B_Mericskay/FranceCarreaux.html.

https://www.sites.univ-rennes2.fr/mastersigat/Webmapping/Mapboxgl/Extrusion_Occitanie_OK.html.

8. Maître de conférence à l'Université Rennes 2, coresponsable du Master SIGAT (Systèmes d'informations géographiques et analyse des territoires).

DES CARTES SUR LE GÉOPORTAIL OU LE SITE DES STATISTIQUES LOCALES

Pour démocratiser l'accès aux données à des utilisateurs moins avertis, une cartographie a donc été proposée sur le site du Géoportail[9] pour toutes les tailles de carreaux (IGN, 2020).

Pour naviguer sur le territoire et zoomer sur des carreaux très fins, l'infrastructure informatique doit en effet être dimensionnée de façon conséquente pour que l'affichage soit fluide. Le Géoportail offre en outre la possibilité de mobiliser en arrière-plan des cartes carroyées de nombreuses autres couches qui viennent enrichir l'information des données. Il peut s'agir du relief, des voies de communication, ou par exemple des zones de crues de la Seine ***(figure 8)*** que l'on peut ainsi croiser avec la densité de population exposée au risque.

Les données sont également accessibles (à la maille du km²) sur le site internet consacré aux statistiques locales de l'Insee (Insee, 2020b).

Preuve toutefois que des solutions intermédiaires entre la mise à disposition de base de données et la cartographie sont possibles, certains utilisateurs experts ont développé des outils permettant d'explorer ces données, d'en exploiter toute la souplesse, en les sélectionnant, les agrégeant, voire en les téléchargeant uniquement sur leur zone d'intérêt, à l'instar par exemple d'OpenDataSoft Explore (ODS, 2020) ou de France en pixel (Francepixel, 2020).

LES DONNÉES CARROYÉES, UNE LOUPE OFFERTE POUR EXPLORER SON PROPRE TERRITOIRE…

Lorsque les données statistiques sont diffusées sous la forme de tableaux, de graphiques, ou de bases de données, il est difficile pour un utilisateur de confronter sa réalité avec les données qu'il manipule.

Démocratiser l'accès à l'information des données présentées dans des carreaux, c'est offrir la possibilité à n'importe quel utilisateur d'aller voir un endroit qu'il connaît, avec deux risques majeurs.

Le premier risque est que l'information affichée lui donne le sentiment que l'information dévoile son intimité. En effet, spontanément, on peut croire que les informations sur les carreaux peu peuplés sont les vraies données. Pour éviter cette perception erronée, l'Insee et l'IGN (qui a la charge du Géoportail), ont travaillé au signalement des traitements apportés à ces données. Ainsi, sur un carreau de 200 mètres, si le nombre de ménages est inférieur à 11, le carreau est hachuré pour signaler que les données sont imputées. De plus, dans l'info-bulle du carreau figure, à côté des données statistiques, l'avertissement suivant : « *Pour des raisons de confidentialité, ces données ont été modifiées.* ». Enfin, des opérations de communication, avec notamment une vidéo pédagogique (Insee, 2019b), ont été réalisées afin d'expliquer la méthodologie employée pour garantir le secret.

Pour autant, au moment de la diffusion des données, et bien que l'accent ait été mis sur ce point dans la documentation, certains utilisateurs se sont inquiétés de repérer

9. Le Géoportail est le portail national de la connaissance du territoire mis en œuvre par l'Institut géographique national (IGN).

sur la carte des informations qu'ils jugeaient trop fines ou dévoilant des données personnelles. Des réponses leur ont été apportées pour détailler les mesures prises pour assurer la confidentialité.

… MAIS PAS DE FAÇON MICROSCOPIQUE

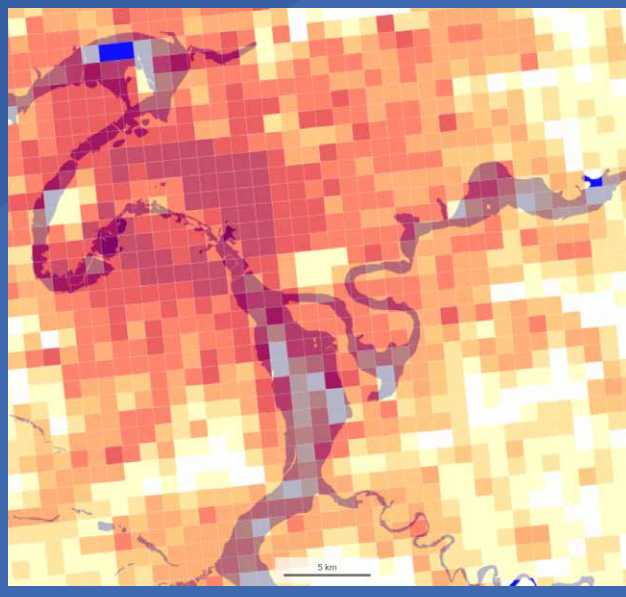
Le second risque est lié au fait que les données carroyées doivent être utilisées pour décrire une zone suffisamment dense, composée de plusieurs carreaux. À cet égard, leur utilité à une maille fine est adaptée à l'analyse urbaine. La valeur affichée sur un seul carreau n'a pas vraiment d'intérêt statistique, mais la finesse de l'information incite souvent l'utilisateur à s'intéresser à celle-ci, sur un endroit familier. Or, les données carroyées souffrent toujours d'une imprécision, en grande partie liée à la localisation de l'information.

En zone rurale, l'exemple classique est celui des grandes parcelles cadastrales comprenant une habitation au milieu d'un champ ou à proximité d'une forêt. De grande taille, la parcelle va être recouverte par plusieurs carreaux, mais l'information statistique ne sera localisée que dans un seul d'entre eux. Parfois, ce carreau « habité » se retrouvera à plusieurs centaines de mètres de l'habitation concernée, dans un lac ou une forêt.

En zone urbaine, ce phénomène se produit également. Sur une grande parcelle cadastrale comprenant plusieurs tours d'immeuble, il se peut que l'étiquette cadastrale soit positionnée

Figure 8. Carte des plus hautes eaux connues sur le bassin de la Seine et Densité de population au carreau

Captures d'écran du Géoportail



Données cartographiques : © Planet Observer INSEE BRGM DRIEE Ile-de-France DREAL Haute-Normandie La Seine en Partage.

une année dans un carreau situé sur une partie de la parcelle et l'année suivante sur le carreau voisin toujours dans la même parcelle. Une analyse en évolution montrera une baisse de population importante sur le premier carreau et une hausse de même ampleur sur le carreau voisin.

> *« Interpréter des informations à une échelle aussi fine doit donc se faire avec prudence. »*

Interpréter des informations à une échelle aussi fine doit donc se faire avec prudence : l'intérêt premier de ces données est de permettre l'analyse de zones urbaines denses constituées de plusieurs carreaux.

QUELLES PERSPECTIVES POUR LES DONNÉES CARROYÉES ?

La diffusion des données carroyées réalisée en 2019 ouvre la voie à l'intégration d'autres sources statistiques sur les grilles de carreau. Le processus de production est désormais décrit et documenté. Il faut pouvoir répondre aux besoins des utilisateurs qui souhaitent que les premières thématiques couvertes par le carroyage (logement, répartition par âge de la population) soient élargies à d'autres champs utiles pour les politiques d'aménagement (emploi, transport, environnement…). « *Si les mises à jour des données relatives à la population et ses caractéristiques sont toujours très attendues, les données de l'emploi suscitent les mêmes impatiences voire exigences* », indique ainsi Caroline De Vellis, « *il nous est également difficile, même en agrégeant plusieurs carreaux, d'obtenir des informations précieuses, comme le croisement de variables, réservées alors à des maillages plus conséquents* ». Pour cela, il faut mobiliser de nouvelles sources statistiques, les géolocaliser et traiter la confidentialité.

Encadré 1. Le carroyage des résultats du recensement de population pour le millésime 2021

Un carroyage va être réalisé afin de répondre à la demande européenne de fourniture de données de population sur des carreaux de 1 km², formulée pour la première fois dans le cadre du *Census 2021* (Eurostat, 2019). Pour le recensement français, cela représente deux défis majeurs :

- le premier est de géolocaliser les lieux de résidences dans les communes de moins de 10 000 habitants. Pour cela, plusieurs méthodes peuvent être utilisées. La première consiste à géolocaliser les adresses figurant sur les documents de collecte, mais elle pose des difficultés dans les communes où l'adressage n'est pas utilisé (en zone rurale par exemple). La seconde méthode, appariement dit « probabiliste », fait le lien entre le recensement et les fichiers fiscaux à partir des caractéristiques individuelles des personnes ;
- dans les communes de plus de 10 000 habitants, les logements sont déjà géolocalisés grâce au Répertoire d'immeubles localisés (RIL). Mais le recensement s'y déroule chaque année par sondage. Le second défi consiste donc à réaliser des estimations fiables sur des carreaux dans ces communes, malgré la non-exhaustivité du recensement sur ces territoires. Des méthodes sont en cours d'expertise pour obtenir des résultats de qualité.

Ces deux défis vont être relevés dans le cadre d'un projet financé par Eurostat. Ces travaux effectués, pour le *Census 2021,* auront à terme un impact sur le système de production du recensement. L'objectif est d'aller au-delà de la réponse au règlement européen, en « industrialisant » la géolocalisation du recensement, pour la production nationale de données sur la population et les logements à partir du recensement. Ceci afin de permettre une diffusion pérenne de données carroyées issues du recensement sur le site ***www.insee.fr.***

Parmi ces sources, on peut citer celles sur l'appareil productif, l'emploi salarié, ou encore le recensement de la population ***(encadré 1).***

S'agissant de cette dernière source, les échéances vont être proches puisque Eurostat souhaite que les résultats du recensement millésimés 2021 soient valorisés sur une maille

Encadré 2. Une grille, des données… pour mesurer l'impact sur l'air et la santé dans un quartier lyonnais

Un exemple intéressant de l'utilisation à la fois de la grille carroyée et des données statistiques de l'Insee est celui que l'on retrouve dans l'étude réalisée dans le cadre du projet de ZAC Part Dieu porté par le Grand Lyon, en 2016 (NumTECH, 2016). Ce projet s'accompagnait de la création de nombreux logements, de bureaux et commerces, et de la modification du schéma de la voirie, qui allaient impacter les trafics automobiles sur la zone d'étude.

L'étude avait pour objectif d'examiner l'impact sur l'air et la santé des riverains. Elle a mobilisé des données carroyées produites par l'Insee (population, figure du haut ci-contre) mais également la grille pour offrir une représentation carroyée d'un indice Pollution/Population (IPP, figure du bas ci-contre), calculé dans cette étude afin d'évaluer l'avant / après du projet. Le calcul de cet indicateur « *repose sur le croisement d'une donnée de pollution (concentration polluante) avec une donnée de population sur le domaine d'étude.* […] *À chaque maille Insee, est affectée la concentration en polluant calculée et la population correspondante. Le calcul de l'IPP est ensuite réalisé en croisant la valeur de population et la concentration. Le résultat fournit un indicateur « d'exposition » de la population.* […] *L'indice a ainsi été évalué pour chaque maille de 200 m de côté de la base Insee (désignées par la suite comme les « mailles Insee »).* »

Ici, **la grille comme les données ont servi de support pour l'aide à la décision.** Les calculs n'auraient pas pu être modélisés à l'échelle de l'IRIS dont la surface est bien trop importante au regard de la problématique des émissions de polluant sur les voiries. Le carreau apporte ici une maille d'analyse indispensable.

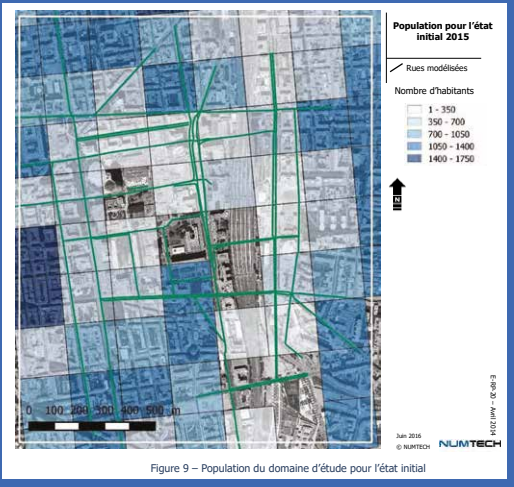


Figure 9 – Population du domaine d'étude pour l'état initial

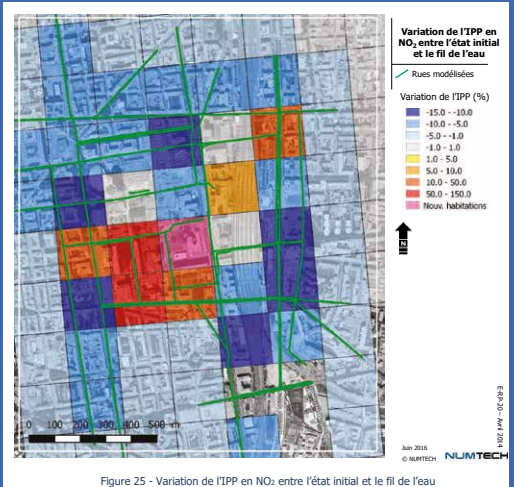


Figure 25 - Variation de l'IPP en NO_2 entre l'état initial et le fil de l'eau

Source : Grand Lyon – Projet ZAC Part Dieu – SETEC environnement Étude Air et Santé (NumTECH, 2016).

carroyée de 1 km² à l'échelle de l'Union européenne (Eurostat, 2019). L'organisme européen met en avant que cette maille de diffusion permettra de mieux répondre aux attentes en « *perpétuelle évolution des utilisateurs, qui attribuent une importance croissante à la disponibilité de données détaillées au niveau local. Cela permettra des analyses beaucoup plus flexibles, même au niveau transfrontalier, adaptables en fonction des besoins politiques et de recherche* ».

Entre injonction européenne et besoins locaux, les demandes en données carroyées pour les études et analyses territoriales sont croissantes. L'Insee a engagé déjà un pas significatif avec la diffusion à l'été 2019 des données carroyées Filosofi 2015. Cette expérience va lui permettre de poursuivre le chemin vers une diffusion plus systématique de données sur cette nouvelle maille géographique. Entre les bases de données pour les utilisateurs avertis et la visualisation en *open data*, des progrès restent à faire pour offrir des fonctionnalités intermédiaires et permettre d'exploiter au maximum la souplesse offerte par cette mosaïque d'informations localisées.

BIBLIOGRAPHIE

APUR, 2014. *Observatoire des quartiers de gare du Grand Paris – Monographie du quartier de gare Les Ardoines – Ligne 15 Sud.* [en ligne]. Juillet 2014. Atelier parisien d'urbanisme. P.6. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *http://www.apur.org/sites/default/files/documents/mononographie_gare_ardoines.pdf*.

BRANCHU, Marc, COSTEMALLE, Vianney et FONTAINE, Maëlle, 2018. Données carroyées et confidentialité. In : *13èmes Journées de Méthodologies Statistiques*. [en ligne]. 12-14 juin 2018. Insee. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *http://www.jms-insee.fr/2018/S23_3_ACTE_BRANCHU_JMS2018.pdf*.

CERTU et CETE NORMANDIE CENTRE, 2011. *Traitements géomatiques par carreaux pour l'observation des territoires.* [en ligne]. Octobre 2011. Éditions du Certu, Collection Dossiers. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *https://www.cerema.fr/fr/centre-ressources/boutique/traitements-geomatiques-carreaux-observation-territoires*.

CNIG, 2020. INSPIRE – Présentation. In : *site du Conseil national de l'information géographique.* [en ligne]. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *http://cnig.gouv.fr/?page_id=8991*.

COSTEMALLE, Vianney, 2018. Identification des problèmes de différenciation géographique à l'aide de la théorie des graphes. In : *13èmes Journées de Méthodologies Statistiques.* [en ligne]. 12-14 juin 2018. Insee. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *http://jms-insee.fr/jms2018s23_2/*.

DE BELLEFON, Marie-Pierre, EUSEBIO, Pascal, FOREST, Jocelyn, PÉGAZ-BLANC, Olivier et WARNOD, Raymond, 2020. *En France, neuf personnes sur dix vivent dans l'aire d'attraction d'une ville.* [en ligne]. 21 octobre 2020. Insee Focus, n° 211. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4806694*.

DELAHAYE, Christine, 1987. Le carroyage : création d'une entité stable. In : *L'Espace géographique.* [en ligne]. Tome 16, n°4, pp. 265-268. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *https://www.persee.fr/doc/spgeo_0046-2497_1987_num_16_4_4270*.

EUROPEAN COMMISSION, 2010. *INSPIRE – Infrastructure for Spatial Information in Europe – D2.8.III.1_v3.0 Data Specification on Statistical Units – Technical Guidelines.* [en ligne]. 10 octobre 2013. European Commission Joint Research Centre. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *http://www.cnig.gouv.fr/wp-content/uploads/2015/01/INSPIRE_DataSpecification_SU_v3.0.pdf*.

EUROSTAT, 2019. *EU legislation on the 2021 population and housing censuses, explanatory notes.* [en ligne]. Février 2019. Theme Population and social conditions, Collection Manuals and guidelines. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/documents/3859598/9670557/KS-GQ-18-010-EN-N.pdf/c3df7fcb-f134-4398-94c8-4be0b7ec0494*.

FLOCH, Jean-Michel, 2012. *Détection des disparités socio-économiques, l'apport de la statistique spatiale.* [en ligne]. 6 décembre 2012. Insee, Direction de la Diffusion et de l'Action régionale, Document de travail N° H2012/04. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/1381028*.

FRANCEPIXEL, 2020. *Site de la France en pixel.* [en ligne]. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *https://www.comeetie.fr/galerie/francepixels2019/#5.7/47/2.3*.

IGN, 2020. *Site du géoportail.* [en ligne]. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *https://www.geoportail.gouv.fr/*.

INSEE, 2019a. *Documentation – données carroyées FILOSOFI 2015.* [en ligne]. Juin 2019. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4176290/documentation_DonneesCarroyees.pdf*.

INSEE, 2019b. *Les données carroyées de l'Insee.* [en ligne]. 27 juin 2019. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4181738*.

INSEE, 2020a. *Production et diffusion des données carroyées.* [en ligne]. 24 février 2020. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/information/3897383*.

INSEE, 2020b. *Statistiques locales.* [en ligne]. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *www.statistiques-locales.insee.fr*.

LAJOIE, Gilles, 1992. *Le Carroyage des informations urbaines – Une nouvelle forme de banque de données sur l'environnement du Grand Rouen.* [en ligne]. Août 2018. Presses universitaires de Rouen et du Havre, nouvelle édition sur OpenEdition Books. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *https://books.openedition.org/purh/8179?lang=fr*.

LOONIS, Vincent et DE BELLEFON, Marie-Pierre, 2018. *Manuel d'analyse spatiale – Théorie et mise en œuvre pratique avec R*. [en ligne]. 29 octobre 2018. Insee, Eurostat, Collection Insee Méthodes, N°131. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/information/3635442*.

NUMTECH, 2016. *Projet PEM / Two Lyon et ZAC Part-Dieu Ouest – Étude air et santé.* [en ligne]. Août 2016. Rapport d'étude pour SETEC Environnement, Réf. 284.1015/ETR – v2.1. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *https://www.grandlyon.com/fileadmin/user_upload/media/pdf/grands-projets/concertation-reglementaire/20170131_gl_zacpartdieu_etudeimpact_2c-etudeairsante.pdf*.

ODS, 2020. *Population française : Données Carroyées à 200 mètres – 2015*. [en ligne]. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *https://public.opendatasoft.com/explore/dataset/population-francaise-donnees-carroyees-a-200-metres-2015/table/*.

OPCS, 1980. *People in Britain: a census atlas.* 1er novembre 1980. Office of Population Censuses and Surveys. Stationery Office Books. ISBN 978-0116906182.

RÉFÉRENCES JURIDIQUES

Directive 2007/2/CE du Parlement européen et du Conseil du 14 mars 2007 établissant une infrastructure d'information géographique dans la Communauté européenne (INSPIRE). In : *Journal officiel de l'Union européenne.* [en ligne]. [Consulté le 3 décembre 2020]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=OJ:L:2007:108:TOC*.

INDICATEURS DE VALEUR AJOUTÉE DES LYCÉES

DU PILOTAGE INTERNE À LA DIFFUSION GRAND PUBLIC

*Franck Evain**

Les indicateurs de valeur ajoutée des lycées (IVAL) mesurent la capacité des établissements à accompagner leurs élèves jusqu'à l'obtention du baccalauréat. Ils sont diffusés depuis 1993 auprès du grand public, par la direction de l'évaluation, de la prospective et de la performance (DEPP) du ministère de l'Éducation. Au-delà du seul taux de réussite à l'examen, les « valeurs ajoutées » associées aux indicateurs bruts facilitent les comparaisons entre des établissements hétérogènes, en prenant en compte les disparités scolaires et socio-économiques entre lycées. Depuis la création des IVAL, à principe inchangé, la méthodologie a évolué, s'adaptant aux enjeux institutionnels et aux données disponibles. Bien que sophistiquée, elle a été choisie de manière à être compréhensible du grand public.
La sortie des indicateurs est un évènement médiatique qui contribue également à leur diffusion éclairée. Les médias, qui produisaient auparavant des classements sur le seul taux de réussite brut, s'orientent désormais davantage vers des approches plurielles. À force de pédagogie, la philosophie des IVAL est de mieux en mieux comprise, au sein du ministère comme à l'extérieur. Si les acteurs institutionnels s'en emparent déjà à des fins de pilotage, cela devrait encore plus être le cas à l'avenir, dans une optique d'évaluation globale des lycées.

In France, the indicators of the added value of lycées (IVAL) measure the ability of secondary schools to support their pupils until they obtain the baccalaureate (A-level). Since 1993 the Ministry of Education's Department of Evaluation, Forward Studies and Performance (DEPP) disseminate it to the general public. In addition to the examination pass rate alone, the 'added values' associated with the raw indicators facilitate comparisons between heterogeneous schools, by taking account of educational and socio-economic disparities between schools. Since the creation of the IVALs, in principle unchanged, the methodology has evolved, adapting to institutional challenges and available data. Although sophisticated, it has been chosen in such a way as to be understandable to the general public.
The release of the indicators is a media event that also contributes to their informed dissemination. The media, which used to produce rankings based solely on the gross success rate, are now moving more towards plural approaches. As a result of this pedagogy, the philosophy of the IVALs is becoming better and better understood, both within the Ministry and externally. Although institutional players are already using them for steering purposes, this should be even more the case in the future, with a view to the overall evaluation of secondary schools.

* Chargé d'étude au Bureau des études sur les établissements et l'éducation prioritaire, DEPP, *franck.evain@education.gouv.fr*

À l'ère de l'*open data,* la diffusion d'indicateurs auprès du grand public semble aller de soi. Pour autant, ce choix était loin d'être aussi évident il y a quelques décennies. Les indicateurs de valeur ajoutée des lycées (IVAL) n'étaient à leur création qu'un outil de pilotage interne, mais leur rôle a rapidement été élargi dans un objectif d'éclairage du débat public. Cette « opération transparence » a fait émerger plusieurs interrogations. Quels indicateurs diffuser ? Par quel biais ? Quelle méthode de calcul utiliser ? Comment éviter les erreurs d'interprétation ? Le statisticien est en effet potentiellement confronté au risque que les journalistes utilisent mal les données qu'il produit. Il lui faut alors arbitrer constamment entre la pertinence statistique des indicateurs diffusés et leur facilité de compréhension par un public non initié. Calculer plusieurs indicateurs pour dépasser une vision monolithique de la performance d'un lycée, prendre en compte les disparités sociales entre établissements, ou encore proposer des grilles de lecture pertinentes, sont autant d'éléments qui ont accompagné la mise en œuvre des IVAL.

LA PERFORMANCE DES LYCÉES, UN QUESTIONNEMENT QUI TRAVERSE LA SOCIÉTÉ

En 1981, *Le Monde de l'éducation* frappe les esprits en publiant le premier palmarès des lycées, établi à partir des seuls taux de réussite bruts au baccalauréat. Le numéro étant un succès à la vente, l'exercice sera reproduit chaque année. Lors de sa création en 1987 au sein du ministère de l'Éducation, la direction de l'Évaluation et de la Prospective (DEP[1]) s'empare du sujet et calcule des indicateurs de performance pour chaque lycée de France[2]. Dès le départ, ces indicateurs ne se limitent pas à des taux bruts, mais prennent en compte les inégalités de recrutement entre établissements. À ce stade, ces données ne sont envoyées qu'aux seuls chefs d'établissement, à des fins de pilotage interne au ministère.

Mais tout change en 1993, lorsque *L'Express* parvient à se procurer un listing incomplet des indicateurs. Le journal en extrait certaines données et intitule son numéro du mois de mars : « *Le classement secret du Ministère de l'Éducation* ». Le ministère apporte alors la réponse suivante : « *Contrairement à ce que vous indiquez, ces indicateurs ne sont pas secrets, puisqu'ils ont été diffusés il y a un an à chaque recteur* [...], *afin qu'il les mette à disposition des chefs d'établissement et s'en serve comme outils d'animation et de pilotage.* [...] *Les indicateurs que vous avez publiés, retenus seuls, tronquent et biaisent la réalité de chaque lycée.* [...] *Que vous ayez réussi à vous les procurer et que vous les ayez diffusés montrent à quel point, à notre regret, vous n'avez pas saisi ce que devaient être les outils d'évaluation et de pilotage, nécessaires par ailleurs, des établissements du second degré ».*

1. Direction de l'Évaluation et de la Prospective, aujourd'hui Direction de l'évaluation, de la prospective et de la performance (DEPP).
2. « *On peut mettre au crédit de Claude Thélot* [directeur de 1990 à 1997], *la mise en place des évaluations diagnostiques de masse, le développement d'indicateurs de performances des lycées, l'enrichissement des publications et du débat sur l'École* [...] » (Cytermann, 2005).

L'ORIGINE DES IVAL : UNE DÉMARCHE DE TRANSPARENCE

Le ministère prend alors les devants. La DEP diffuse, l'année suivante, la première version des indicateurs de valeur ajoutée des lycées (IVAL), obtenus à partir des résultats de la session 1993 du baccalauréat. L'objectif est d'opposer au « sensationnalisme » journalistique la transparence d'une démarche scientifique (Buisson-Fenet, 2019). Les indicateurs fournis par la DEP sont repris en juin 1994 par plusieurs journaux, dont *L'Express,* qui débute son article par *« Finie la politique du silence ! »*. L'hebdomadaire indique ensuite avoir reçu de nombreux courriers de proviseurs se plaignant de son précédent classement, qui ne tenait pas suffisamment compte du contexte. *L'Express* fait amende honorable en précisant : *« Le taux de réussite au bac* [...] *doit être comparé au taux de réussite attendu, c'est-à-dire au résultat que devrait obtenir le lycée, compte tenu de l'âge des élèves et de leur origine sociale. Cette donnée, calculée par ordinateur, permet de découvrir la valeur ajoutée – positive ou négative – de l'établissement ».*

> *« L'objectif est d'opposer au « sensationnalisme » journalistique la transparence d'une démarche scientifique. »*

Si la diffusion grand public des indicateurs ne s'est pas faite de manière paisible et linéaire, cela n'a toutefois pas empêché leur installation durable dans le paysage médiatique. Ainsi, les années suivantes, la plupart des journaux et hebdomadaires emboîtent le pas, rivalisant de titres accrocheurs : *« Les meilleurs lycées ne sont pas ceux qu'on croit » (Le Nouvel Observateur,* 1997*), « La vérité sur les bons et les mauvais lycées » (L'Express,* 1998*), « Lycées : les effets pervers du palmarès. Ils reposent sur des indicateurs réducteurs » (Libération,* 1999*), etc.*

Un long chemin a été parcouru depuis 1993. Les nombreuses réformes portant sur l'organisation des lycées n'ont jamais remis en cause l'existence du baccalauréat, élément central de l'évaluation dans leur performance ***(encadré 1)***. Mais en 27 ans d'existence, la nature, le champ et la méthode de calcul des IVAL ont été l'objet d'évolutions constantes.

QUELS INDICATEURS, POUR QUELLE UTILISATION ?

Trois indicateurs sont utilisés pour mesurer la capacité des lycées à accompagner leurs élèves jusqu'au baccalauréat, dont le point de départ est pour chacun d'entre eux un **taux brut** ou **observé** :

- Le premier porte sur la réussite au baccalauréat**.** Le **taux de réussite** rapporte le nombre d'élèves du lycée reçus à l'examen au nombre d'élèves qui s'y sont présentés. C'est un indicateur traditionnel, connu et facile à établir.
- Le **taux d'accès** évalue la probabilité, pour un élève, d'obtenir le baccalauréat à l'issue d'une scolarité entièrement effectuée dans le lycée, même s'il y a redoublé. Le taux d'accès de la seconde au baccalauréat est le produit de trois taux intermédiaires : de la seconde à la première, de la première à la terminale et de la terminale au baccalauréat.
- Enfin, le troisième indicateur concerne le **taux de mentions** au baccalauréat, qu'il s'agisse d'une mention « Assez bien », « Bien » ou « Très bien ».

Pour comparer des lycées entre eux, il est nécessaire de tenir compte des caractéristiques des élèves qu'ils accueillent et de leur offre de formation : si un lycée présente une valeur élevée pour un indicateur, est-ce dû au fait qu'il a reçu des élèves ayant un très bon niveau scolaire, ou au fait qu'il a su, tout au long de la scolarité, développer chez eux les connaissances et les méthodes de travail qui ont permis leur succès ?

Encadré 1. L'enseignement du second degré en France

L'enseignement du second degré est dispensé dans les collèges, puis dans les lycées, généraux et technologiques ou professionnels.

Les élèves ont en moyenne 11 ans lorsqu'ils quittent l'école pour entrer au collège. Ils y restent quatre années, à l'issue desquelles ils passent l'examen du Diplôme National du Brevet (DNB), qui atteste de leur acquisition de connaissances générales.

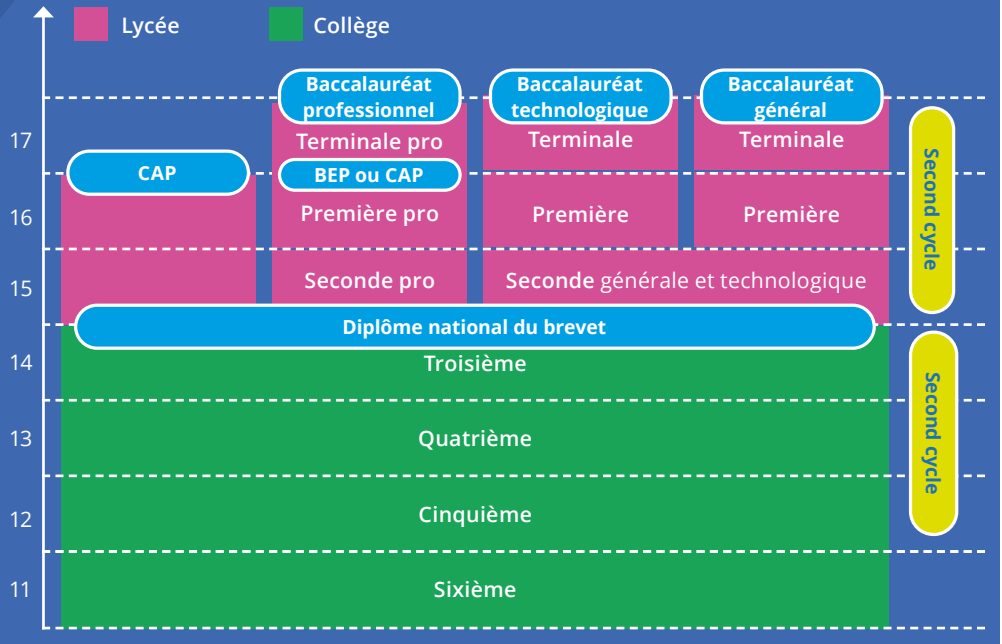
Chaque année, 800 000 élèves passent le brevet. Les trois quarts d'entre eux s'orientent ensuite dans des lycées généraux et technologiques, et un quart dans des lycées professionnels.

En point d'orgue de trois années d'enseignement en lycée, les élèves passent enfin le diplôme du baccalauréat, sanctionnant la fin des études secondaires. Il existe trois types de baccalauréat :

- Baccalauréat général, où les nouveaux « enseignements de spécialité », introduits en 2019, remplacent désormais les traditionnelles séries littéraire, économique et social et scientifique.
- Baccalauréat technologique : sept séries, dont « Sciences et technologies du management et de la gestion » (STMG), « Sciences et technologies de l'industrie et du développement durable » (STI2D), etc.
- Baccalauréat professionnel : de nombreuses séries, regroupées pour les besoins des IVAL en dix « domaines de spécialités » (« Mécanique, électricité, électronique », « Communication et information », « Services aux personnes », etc.).

Certains lycées, dits « polyvalents », accueillent à la fois des élèves de la voie générale et technologique et des élèves de la voie professionnelle. Dans les IVAL, ces deux voies sont alors traitées à part.

Lors de la session 2019 du baccalauréat, 390 000 candidats se sont présentés dans la voie générale, 156 000 dans la voie technologique et 209 000 dans la voie professionnelle.



Pour chaque taux observé est ainsi calculé un **taux attendu** ou **prédit**. Ils correspondent aux taux moyens des lycées accueillant des élèves aux caractéristiques identiques.

La **valeur ajoutée** d'un indicateur est l'écart entre le taux observé et le taux attendu. Elle évalue l'apport propre de l'établissement, compte tenu du profil initial de ses élèves ***(figure 1)***.

Ces indicateurs sont *complémentaires*, car ils prennent non seulement en compte la réussite à l'examen final, mais aussi la capacité des lycées à accompagner leurs élèves depuis leur entrée dans le lycée jusqu'à l'obtention du baccalauréat. En effet, que signifierait un très bon taux de réussite dans un établissement où seulement la moitié des élèves entrés dans le lycée seraient encore présents en terminale ?

> *« La valeur ajoutée évalue l'apport propre de l'établissement, compte tenu du profil initial de ses élèves. »*

De manière générale, tout indicateur statistique donne une vision restrictive de la réalité. Même des indicateurs abondamment utilisés, comme le produit intérieur brut (PIB) en économie, sont accusés d'être incomplets et d'écarter des domaines primordiaux tels que la qualité de vie ou le développement durable[3].

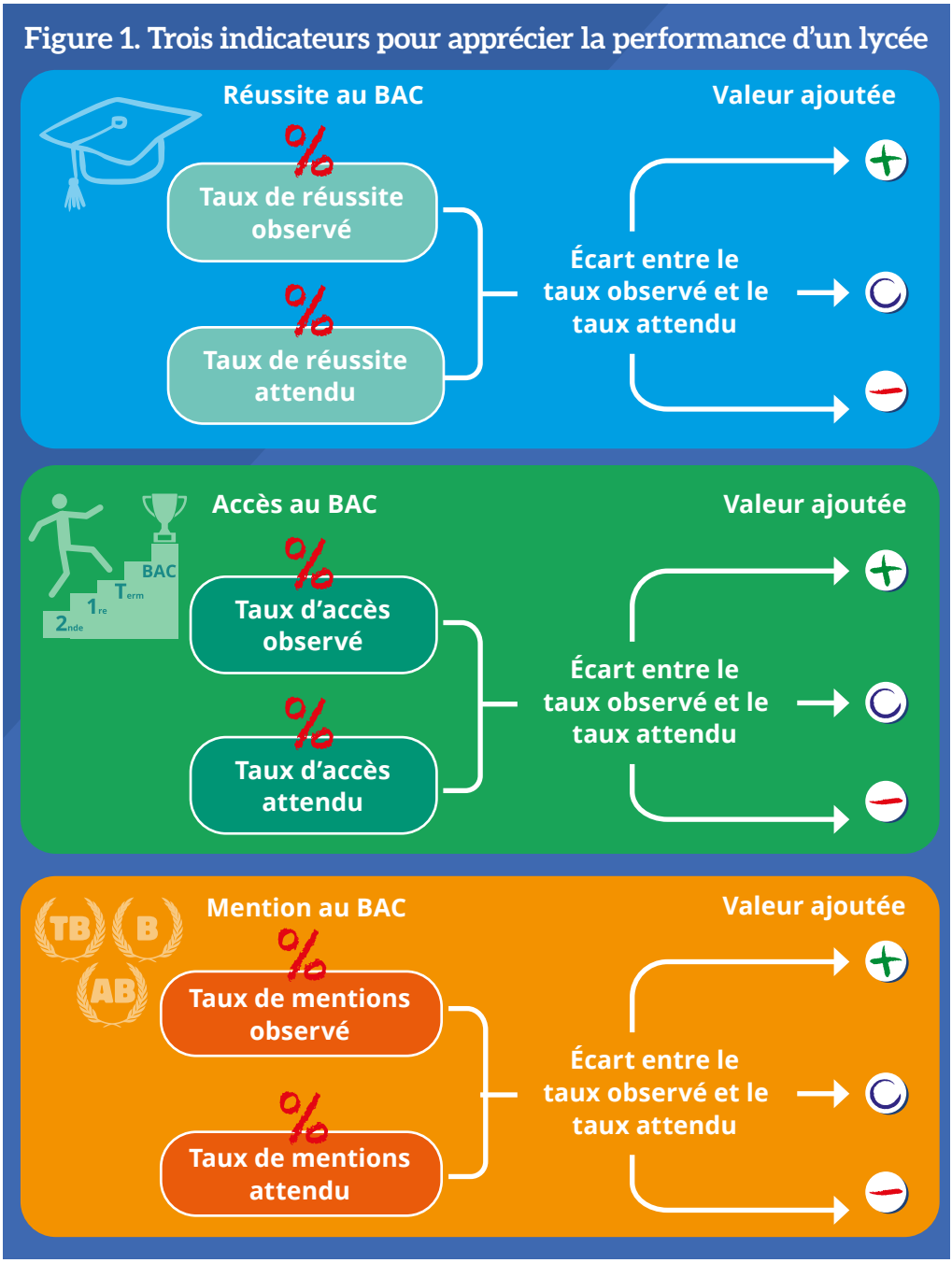
Dans le cas de la mesure de la performance des lycées, le taux d'accès répond à la problématique de parcours, en prenant en compte l'ensemble de la scolarité dans l'établissement. Associés, les trois indicateurs fournissent une vision plus juste et plus complète de l'action du lycée. Leur nombre réduit facilite par ailleurs leur lecture et les éventuelles reprises dans le débat public.

Chaque année, au mois de mars, la DEPP met à disposition du grand public, pour tous les lycées généraux, technologiques et professionnels, publics et privés, les taux observés des trois indicateurs et les valeurs ajoutées qui y sont associées ***(figure 2)***. Les taux de réussite et de mentions sont déclinés par série, et le taux d'accès de la seconde au baccalauréat est complété par les taux première-bac et terminale-bac. Ce dernier permet notamment de mesurer la part de redoublants acceptés au sein de l'établissement.

Ces indicateurs servent en premier lieu à éclairer le débat public sur le système éducatif, dans une démarche de transparence. Mais pas seulement. Ce sont également des outils de pilotage à destination des chefs d'établissement, des inspecteurs d'académie et des recteurs. Ils leur permettent d'apprécier les résultats des lycées dont ils ont la responsabilité. À un niveau plus fin, les valeurs ajoutées de chaque série peuvent aussi être des éléments de réflexion pour les équipes éducatives, et notamment les enseignants. Enfin, les IVAL sont des éléments d'information, parmi beaucoup d'autres, pour les parents d'élèves. Ils permettent de dépasser une interprétation forcément simpliste des seuls taux de réussite observés, et contribuent à enrichir le dialogue entre les parents et les établissements.

3. En 2015, le Conseil économique, social et environnemental a d'ailleurs proposé dix indicateurs complémentaires au PIB pour prendre en compte toutes les dimensions du développement, tant économiques et sociales qu'environnementales.

Figure 1. Trois indicateurs pour apprécier la performance d'un lycée



COMMENT DÉTERMINER LA VALEUR AJOUTÉE ?

Lorsque des indicateurs sont publiés sur un grand nombre d'individus, au sens statistique du terme, la question de la comparabilité est majeure.

Plusieurs approches sont alors possibles, par exemple celle de la Haute Autorité de Santé (HAS) lorsqu'elle publie des indicateurs de qualité des soins pour chaque hôpital et clinique de France[4]. La problématique est proche de celle des IVAL, mais le traitement sensiblement différent. Sur le site public de la HAS, chaque établissement se voit affecter une note (A, B, C, D ou E) sur une batterie d'indicateurs. Mais un « A » ne signifie pas la même chose si 80 % des établissements sont dans ce cas, ou si 80 % ont obtenu une moins bonne note. Le site mentionne alors pour chaque indicateur la distribution des notes au niveau national et positionne l'hôpital ou la clinique par rapport à ses semblables. Il offre également la possibilité de comparer plusieurs établissements sur chacun des critères d'évaluation.

> *« Afin d'expliquer au mieux l'impact d'un lycée sur la réussite de ses élèves, il faut ainsi s'efforcer d'éliminer l'incidence des facteurs de réussite qui lui sont extérieurs. »*

Dans le cas particulier des lycées, les indicateurs portent sur quelque 4 300 établissements, aux caractéristiques très différentes les uns des autres, notamment en termes de profil des élèves accueillis. Afin d'expliquer au mieux l'impact d'un lycée sur la réussite de ses élèves, il faut ainsi s'efforcer d'éliminer l'incidence des facteurs de réussite qui lui sont extérieurs.

Une partie de ces facteurs est en effet propre à l'élève. Les quatre caractéristiques individuelles que sont l'âge, le sexe, l'origine sociale et le niveau scolaire à l'entrée au lycée ont été retenues, car elles donnent une première approximation des chances d'accès et de réussite au baccalauréat d'un élève. Le facteur qui a l'impact le plus fondamental sur la réussite est le *niveau à l'entrée au lycée* : il est mesuré *via* la note moyenne obtenue aux épreuves écrites du diplôme national du brevet (DNB). L'écart sur la réussite au baccalauréat est de près de 18 points entre les élèves ayant eu 10 ou moins à ces épreuves écrites et ceux ayant eu plus de 14[5]. L'origine sociale des élèves est quant à elle mesurée par *l'indice de position sociale*. Cet indice, propre à l'Éducation Nationale, permet de prendre en compte, à travers une variable continue, les catégories socioprofessionnelles des deux parents : plus l'indice est élevé, plus le milieu social de l'élève est favorisé (Rocher, 2016). Évidemment, ne peuvent être pris en compte que des facteurs mesurables. Le degré d'implication des parents dans la scolarité de leurs enfants, s'il joue forcément un rôle, ne peut par exemple être retenu.

L'autre partie des facteurs de réussite d'un élève est liée aux caractéristiques des élèves qui l'entourent. La plupart des travaux portant sur la mixité concluent à une influence – ou effet d'entraînement – du profil scolaire et socio-économique des camarades de classe d'un élève sur ses résultats scolaires (Fougère, Givord, Monso et Pirus, 2019). Aux quatre caractéristiques individuelles sont ainsi ajoutés leurs équivalents « collectifs » : proportion d'élèves en retard scolaire, proportion de filles, indice de position sociale moyen et note moyenne des élèves aux épreuves écrites du brevet.

4. Pour plus de détail, consulter *https://www.scopesante.fr*.
5. Chiffres obtenus lors de la session 2019 du baccalauréat.

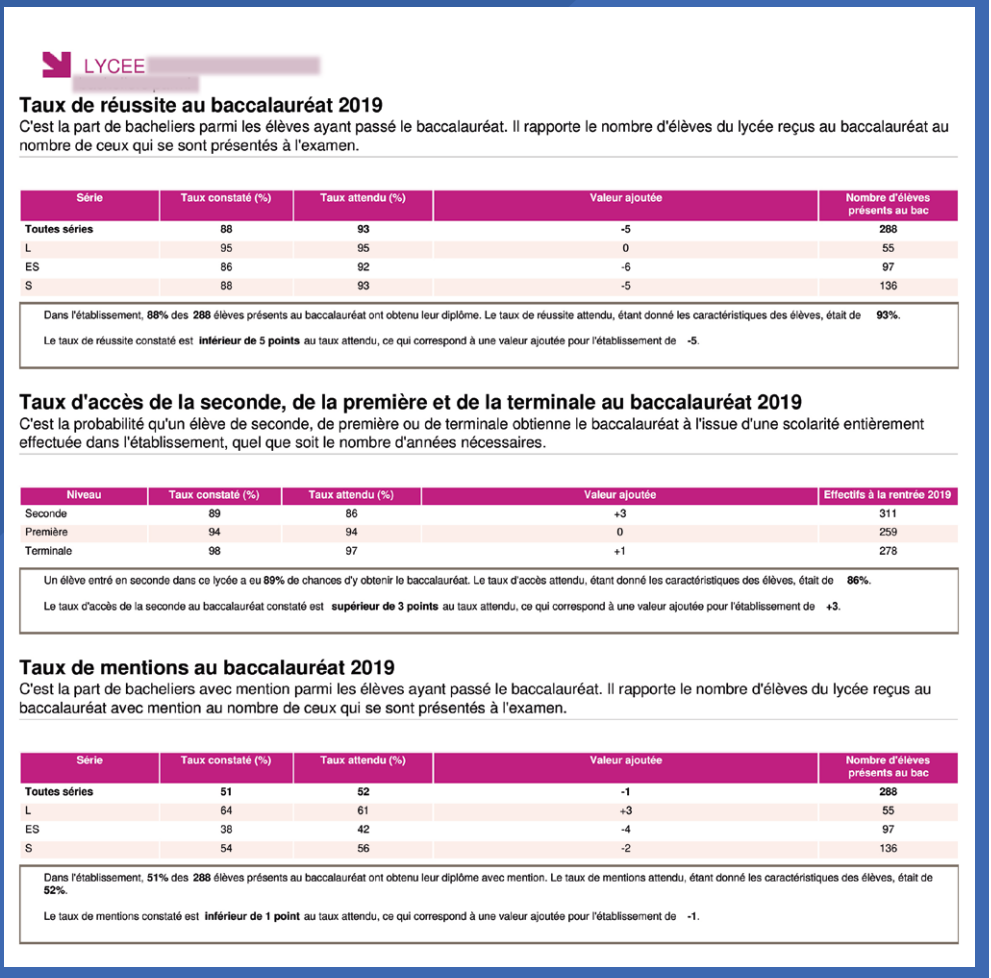
LES MODÈLES MIS EN ŒUVRE DANS LE CADRE DES IVAL

La modélisation vise à expliquer les taux de réussite, d'accès et de mentions par les caractéristiques à la fois individuelles et collectives des élèves. Elle s'appuie sur des modèles logistiques multiniveaux où une observation est un élève ***(encadré 2)***.

Le taux de réussite attendu, pour un élève, s'obtient en appliquant à ses caractéristiques (individuelles et collectives) les valeurs des coefficients du modèle, desquels on retire l'effet établissement.

En procédant de cette manière, le taux attendu (ou prédit) correspond à la probabilité de réussite de l'élève s'il était dans un lycée « moyen », au sens statistique du terme.

Figure 2. Exemple d'une fiche IVAL disponible sur le site du Ministère



LYCEE [redacted]

Taux de réussite au baccalauréat 2019

C'est la part de bacheliers parmi les élèves ayant passé le baccalauréat. Il rapporte le nombre d'élèves du lycée reçus au baccalauréat au nombre de ceux qui se sont présentés à l'examen.

Série	Taux constaté (%)	Taux attendu (%)	Valeur ajoutée	Nombre d'élèves présents au bac
Toutes séries	88	93	-5	288
L	95	95	0	55
ES	86	92	-6	97
S	88	93	-5	136

Dans l'établissement, **88%** des **288** élèves présents au baccalauréat ont obtenu leur diplôme. Le taux de réussite attendu, étant donné les caractéristiques des élèves, était de **93%**.

Le taux de réussite constaté est **inférieur de 5 points** au taux attendu, ce qui correspond à une valeur ajoutée pour l'établissement de **-5**.

Taux d'accès de la seconde, de la première et de la terminale au baccalauréat 2019

C'est la probabilité qu'un élève de seconde, de première ou de terminale obtienne le baccalauréat à l'issue d'une scolarité entièrement effectuée dans l'établissement, quel que soit le nombre d'années nécessaires.

Niveau	Taux constaté (%)	Taux attendu (%)	Valeur ajoutée	Effectifs à la rentrée 2019
Seconde	89	86	+3	311
Première	94	94	0	259
Terminale	98	97	+1	278

Un élève entré en seconde dans ce lycée a eu **89%** de chances d'y obtenir le baccalauréat. Le taux d'accès attendu, étant donné les caractéristiques des élèves, était de **86%**.

Le taux d'accès de la seconde au baccalauréat constaté est **supérieur de 3 points** au taux attendu, ce qui correspond à une valeur ajoutée pour l'établissement de **+3**.

Taux de mentions au baccalauréat 2019

C'est la part de bacheliers avec mention parmi les élèves ayant passé le baccalauréat. Il rapporte le nombre d'élèves du lycée reçus au baccalauréat avec mention au nombre de ceux qui se sont présentés à l'examen.

Série	Taux constaté (%)	Taux attendu (%)	Valeur ajoutée	Nombre d'élèves présents au bac
Toutes séries	51	52	-1	288
L	64	61	+3	55
ES	38	42	-4	97
S	54	56	-2	136

Dans l'établissement, **51%** des **288** élèves présents au baccalauréat ont obtenu leur diplôme avec mention. Le taux de mentions attendu, étant donné les caractéristiques des élèves, était de **52%**.

Le taux de mentions constaté est **inférieur de 1 point** au taux attendu, ce qui correspond à une valeur ajoutée pour l'établissement de **-1**.

(*https://www.education.gouv.fr/les-indicateurs-de-resultats-des-lycees-1118*)

Autrement dit, si l'impact de son établissement sur sa réussite était neutre. Les taux de réussite attendus par lycée ou par série sont obtenus en faisant la moyenne sur l'ensemble des élèves concernés[6].

Le procédé est identique pour les taux de mentions. Pour le taux d'accès, la méthode est reproduite pour chacun des trois taux intermédiaires (seconde-première, première-terminale et terminale-bac) et le taux seconde-bac attendu est obtenu en faisant le produit de ces trois taux ***(figure 3)***.

Ce concept de *taux attendu* est essentiel. Comparer tels quels les taux bruts de deux lycées, sans tenir compte des caractéristiques de leurs élèves, conduirait en effet à une analyse faussée. Si des lycées ont de faibles taux attendus, c'est parce qu'ils accueillent des élèves d'un moins bon niveau scolaire et aux caractéristiques sociales plus défavorisées que les autres. Il serait déloyal, notamment pour les équipes éducatives, de comparer leurs taux bruts à ceux de lycées très favorisés.

Comme indiqué plus haut, la valeur ajoutée est obtenue, pour chacun des indicateurs (réussite, accès et mentions), en comparant taux observé et taux attendu. Elle permet une comparaison avec l'efficacité moyenne, et mesure de manière beaucoup plus juste ce qu'a apporté le lycée à ses élèves.

Cette méthode, appuyée sur une modélisation adaptée, n'a pas toujours été employée pour calculer les IVAL. Elle résulte d'un long cheminement marqué par des évolutions méthodologiques et institutionnelles.

UNE MÉTHODOLOGIE QUI S'EST ADAPTÉE AUX DONNÉES DISPONIBLES…

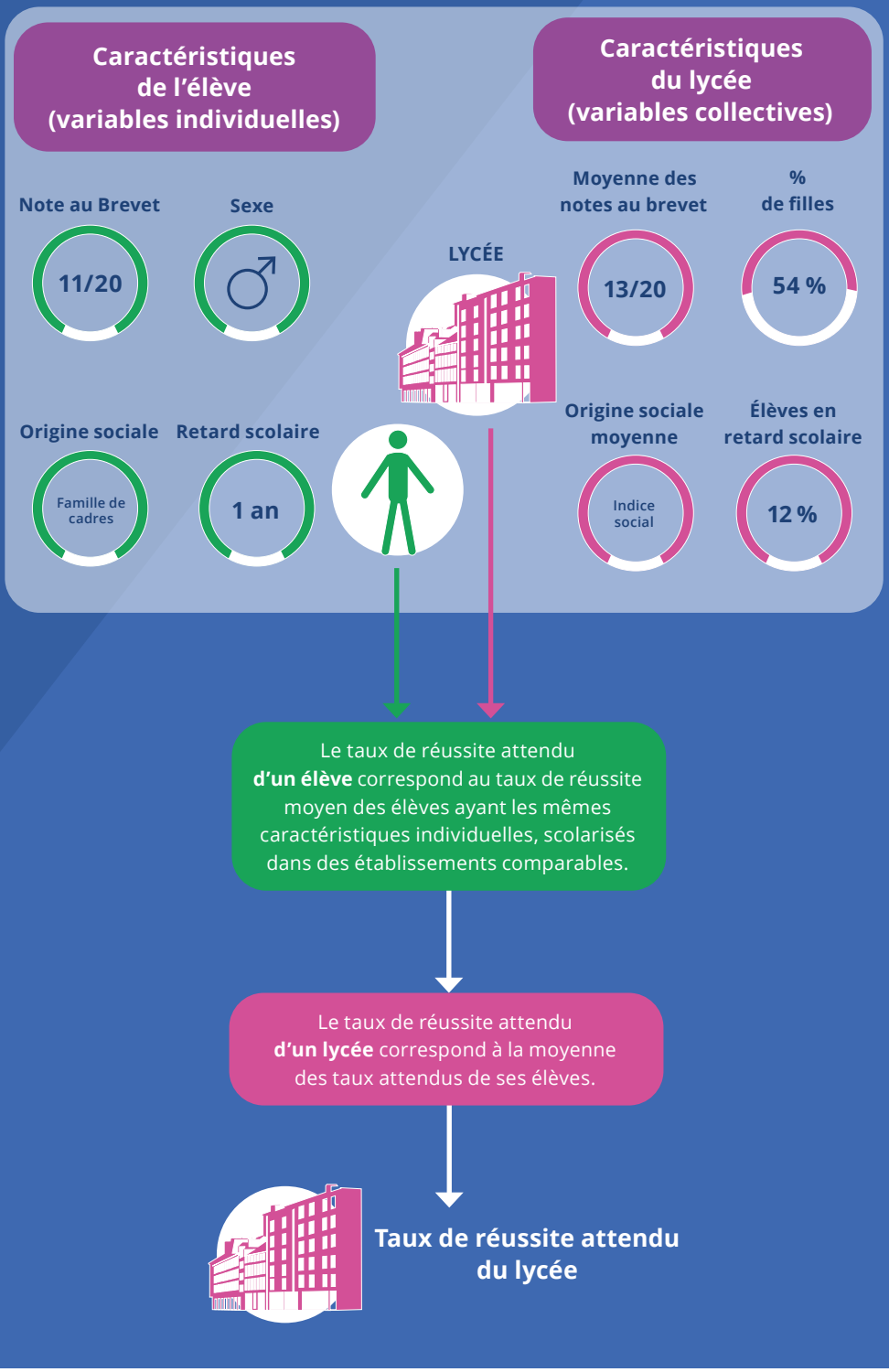
Avant 2008, la manière de calculer le taux attendu était beaucoup plus fruste et ne prenait en compte que la catégorie socioprofessionnelle du responsable légal (en quatre modalités) et l'âge en terminale (en trois modalités). Pour chaque lycée, était calculée la répartition des élèves dans les 12 cases résultant du croisement de ces deux critères. À chaque case était ensuite affecté le taux observé au niveau national pour cette catégorie d'élèves. La moyenne de ces taux de référence, pondérés par la répartition dans les 12 cases, donnait le taux attendu du lycée.

En 2008, les IVAL ont fait l'objet d'une importante refonte (Duclos et Murat, 2014). D'une part, la liste des variables explicatives s'est enrichie, et d'autre part, le calcul des taux attendus s'est appuyé pour la première fois sur des modèles économétriques.

Dans les premières années de diffusion des IVAL, certains observateurs regrettaient en effet que le niveau des élèves à l'entrée au lycée ne fasse pas partie des critères retenus. Des travaux expérimentaux ont pourtant été menés dès 1996, afin d'évaluer l'impact de la prise en compte des notes au brevet. En 2004, dans une étude commandée par la DEPP portant sur les lycées de l'académie de Bordeaux, l'auteur concluait que la non prise en compte du niveau initial des élèves conduisait à une sous-évaluation de la valeur ajoutée des lycées populaires et à une surévaluation de la valeur ajoutée des lycées socialement favorisés (Félouzis, 2004).

6. Pour plus de détails sur la partie modélisation, voir (Evain et Évrard, 2017).

Figure 3. Comment est calculé le taux de réussite «attendu» d'un lycée



Mais les données nécessaires n'étaient alors pas disponibles. Les résultats aux épreuves terminales du brevet étaient encore collectés « à la main » dans bon nombre de départements et ne constituaient pas une base de données utilisable à des fins statistiques. La transformation en 2003 du brevet en un examen national[7] et la remontée de données exhaustives au niveau national à partir de la session 2004 ont permis de remédier au problème. Il a néanmoins fallu attendre quelques années supplémentaires, car les élèves ayant passé le brevet en 2004 ne sont arrivés en terminale, au mieux, que trois ans plus tard.

En outre, si un fichier national du brevet était bien disponible en 2008, des difficultés d'appariement avec les autres sources de données subsistaient. Afin de tenir compte, malgré tout, du niveau initial des élèves, celui-ci est calculé à un niveau agrégé. Pour chacun des élèves de terminale S d'un lycée, par exemple, c'est la moyenne des notes au brevet de l'ensemble des élèves de terminale S de ce lycée qui est utilisée. À cette occasion, le sexe est également ajouté aux variables explicatives. Les quatre dimensions que ces variables permettent de prendre en compte sont restées les mêmes depuis lors : niveau initial des élèves, origine sociale, retard scolaire et sexe.

> *« Sept ans plus tard, en 2015, les possibilités d'appariement des différentes sources se sont nettement améliorées, en lien avec les progrès de l'immatriculation des élèves. »*

La deuxième grande nouveauté de la refonte de 2008 concerne la méthode. Le taux attendu est désormais calculé à l'aide de modèles logistiques multiniveaux, de la manière décrite précédemment ***(encadré 2)***. Ces modèles légitiment l'ajout de variables de contexte, qui apparaissent ainsi pour la première fois : répartition des élèves par PCS, proportion de filles, proportion d'élèves en retard scolaire et note moyenne au brevet.

Sept ans plus tard, en 2015, les possibilités d'appariement des différentes sources se sont nettement améliorées, en lien avec les progrès de l'immatriculation des élèves. Grâce à un appariement sur identifiant individuel crypté, il est alors devenu possible de récupérer la note au brevet de chaque élève. C'est aussi en 2015 que l'indice de position sociale (voir *supra*) a été créé, permettant de prendre en compte la profession des deux parents, autre avancée très importante. Tous ces éléments, sur lesquels le lycée n'a pas de prise, sont ainsi de mieux en mieux mesurés. Cela permet une meilleure estimation des taux attendus, et donc une meilleure comparaison « toutes choses égales par ailleurs » des lycées entre eux.

… ET AU CONTEXTE INSTITUTIONNEL

Ces améliorations successives ont été rendues possibles grâce à l'éclosion de nouveaux outils ou de nouvelles sources de données. D'autres se sont imposées, afin de tenir compte des évolutions du contexte institutionnel. Par exemple, entre 2009 et 2011, le baccalauréat professionnel a été profondément réformé. Alors qu'il s'obtenait auparavant en deux ans, sa durée s'est alignée sur celle du baccalauréat général. Les lycées professionnels accueillent depuis des élèves directement issus du collège, et ce pour une durée de trois ans. Dès lors, les IVAL se sont adaptés pour fournir un taux d'accès de la seconde au baccalauréat pour les lycées professionnels.

7. Auparavant, les épreuves étaient académiques.

Plus récemment, les IVAL se sont enrichis d'un nouvel indicateur, le taux de mentions, pour répondre à une problématique récurrente. Si les taux de réussite se stabilisent depuis 2014, ils n'ont en effet cessé d'augmenter les années précédentes. En 2019, ils atteignaient 90 % dans la voie générale et technologique et 82 % dans la voie professionnelle. Les taux attendus suivent logiquement la même évolution. Pour un lycée général et technologique sur six, le taux de réussite attendu est même supérieur ou égal à 97 %. Par construction, la valeur ajoutée de ces lycées, qui accueillent des élèves au profil très favorisé, ne peut alors dépasser +3 points. Parmi les critiques adressées aux IVAL, la difficulté de discriminer les lycées très favorisés revenait ainsi régulièrement dans le débat. En 2017, l'ajout du taux de mentions a permis d'y apporter une réponse. En effet, pour ces lycées, le taux moyen de mentions est de 75 %, ce qui est certes élevé, mais leur laisse davantage de latitude pour obtenir une bonne valeur ajoutée. Parmi les lycées très favorisés des IVAL 2019, on observe d'ailleurs des situations très contrastées. Si les valeurs ajoutées associées à leur taux de réussite sont globalement proches de zéro, l'un d'entre eux a par exemple une valeur ajoutée de −15 sur le taux de mentions, tandis que pour un autre elle est de +27. L'ajout de ce nouvel indicateur permet ainsi de différencier des lycées qui paraissaient auparavant très proches.

Le taux de mentions constitue par ailleurs un indicateur de pilotage en lui-même. Pour les élèves qui poursuivent leurs études dans le supérieur, le type de parcours et le taux d'obtention du diplôme varient grandement selon que l'élève a eu son baccalauréat avec mention ou non. En licence, le taux de diplômés est de 77 % parmi les élèves ayant obtenu une mention, contre seulement 47 % pour les élèves qui n'en ont pas eu (Ponceau, 2019). Cet indicateur mesure donc également la capacité des lycées à préparer leurs élèves aux études supérieures.

Encore plus récemment, la réforme de la voie générale du baccalauréat, entamée en 2019, a vu les « enseignements de spécialité » remplacer les traditionnelles séries générales. Pour l'année scolaire 2019-2020, seuls les élèves de première étaient concernés. En lieu

Encadré 2. Les modèles utilisés dans les IVAL

Les modèles utilisés sont des modèles logistiques multiniveaux, où une observation égale un élève.

- Logistiques, pour tenir compte de la nature dichotomique de la variable expliquée au niveau élève : réussite ou non, mention ou non, accès au niveau supérieur ou non ;
- Multiniveaux, car ce type de modèle permet de mesurer les effets de contexte sur les individus, dans le cas où ces derniers partagent un environnement commun. Ici, les élèves d'un lycée sont tous soumis au même environnement, mesuré à travers les variables collectives décrites ci-dessus.

Une modélisation multiniveaux permet alors de mieux estimer l'effet des variables individuelles et collectives auxquelles on s'intéresse (Givord et Guillerm, 2016). Une régression classique, par les moindres carrés ordinaires, aurait de plus supposé une indépendance des termes d'erreur d'un individu à l'autre. Cette hypothèse est ici mise à mal : on suppose au contraire que l'environnement a un impact sensiblement équivalent sur tous les individus d'un même groupe. Or, la non indépendance des termes d'erreur peut conduire à une mauvaise estimation des coefficients. Une modélisation multiniveaux permet de résoudre ce problème, en décomposant le terme d'erreur du modèle en un terme strictement individuel, et un terme commun à tous les individus d'un même groupe. Ce second terme correspond à l'effet de contexte, c'est-à-dire, dans le cadre des IVAL, à l'effet établissement, qui prend une valeur identique pour tous les élèves d'un même établissement.

et place des séries littéraire, économique et sociale ou scientifique, ils avaient à choisir trois spécialités parmi treize[8]. Or, les taux de réussite et de mentions sont actuellement calculés par série. Lorsque ces élèves passeront le baccalauréat, en juin 2021, il sera alors nécessaire d'adapter les indicateurs pour tenir compte de la disparition des séries générales.

Ces ajustements ne sont que quelques exemples parmi d'autres. Les réformes touchant au lycée, les critiques constructives des proviseurs ou la recherche d'une meilleure pertinence des indicateurs nécessitent en effet de repenser chaque année la méthodologie, afin qu'elle reste la plus appropriée possible.

Ces différentes adaptations font bien sûr l'objet de nombreux travaux en amont. Un important travail de communication ***(figure 4)*** est également effectué en aval, à destination des utilisateurs des IVAL et, en premier lieu, du grand public.

DIFFUSER DES INDICATEURS ET LES EXPLIQUER

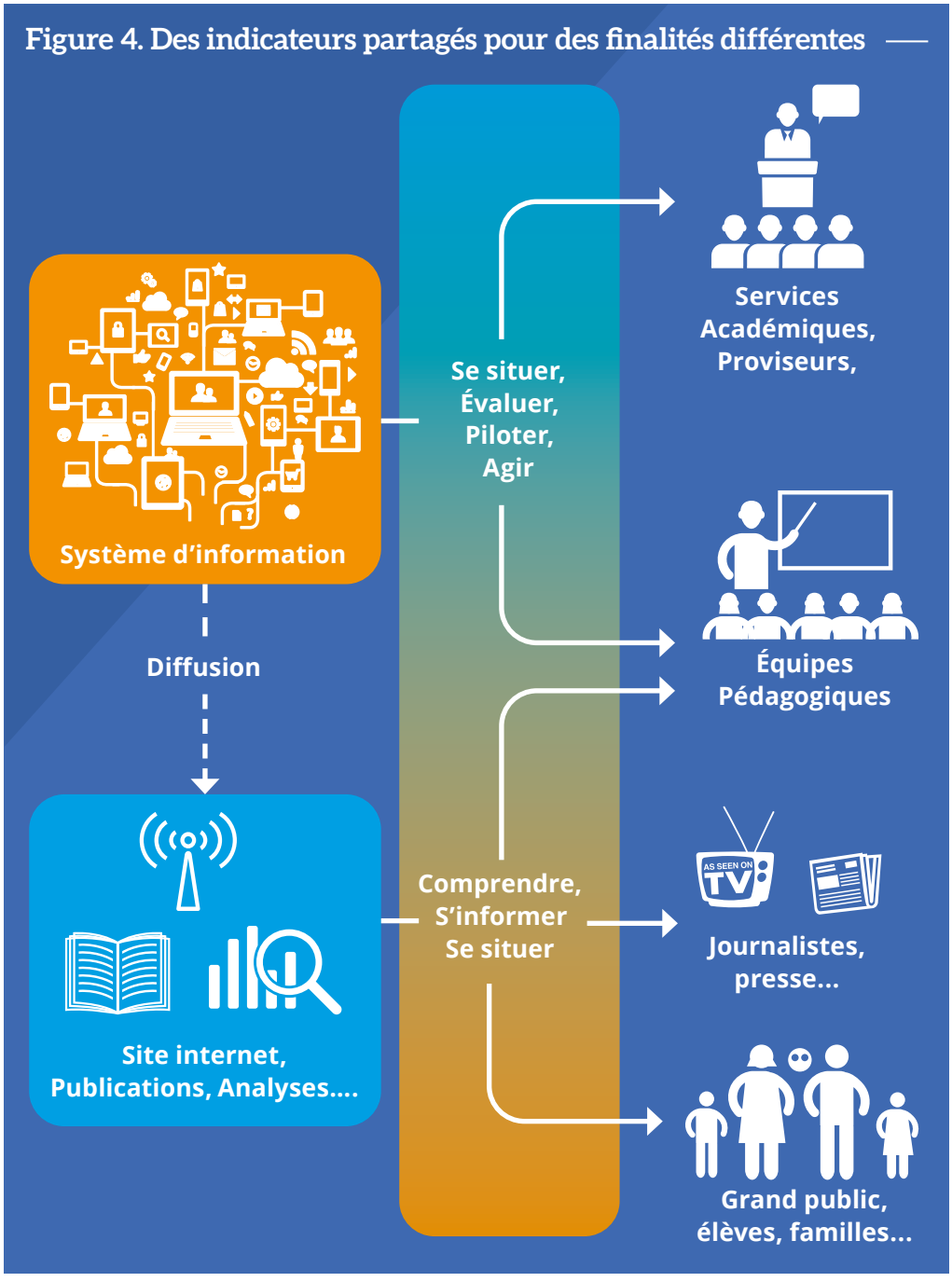
Le critère de lisibilité et de compréhension est déterminant pour un public non statisticien et potentiellement rétif aux chiffres. La manière dont est calculée la valeur ajoutée illustre d'ailleurs ce souci d'être compris du plus grand nombre. En effet, d'autres choix que celui de mesurer l'écart entre taux observé et taux attendu auraient pu être faits. Par exemple, puisque le modèle permet de mettre en évidence un effet propre à chaque établissement, ce coefficient aurait pu être considéré comme correspondant à la valeur ajoutée. Cependant, sa grandeur (en valeur absolue) ne correspond à aucune échelle. Sauf à l'utiliser uniquement pour comparer deux lycées entre eux et voir lequel a la meilleure valeur ajoutée, ce coefficient aurait donc été difficile à interpréter. Le choix qui a été retenu permet d'expliquer au grand public, dans des termes relativement simples, ce à quoi correspond la valeur ajoutée : l'écart entre l'observé et ce à quoi on pouvait s'attendre, compte tenu des caractéristiques des élèves.

> *« Le choix qui a été retenu permet d'expliquer au grand public, dans des termes relativement simples, ce à quoi correspond la valeur ajoutée. »*

Afin d'expliquer les grands principes et la méthodologie des indicateurs, la diffusion des résultats s'accompagne d'un certain nombre d'éléments à visée pédagogique. Tous les ans, la DEPP organise ainsi une conférence de presse, au cours de laquelle sont présentés les indicateurs, leurs objectifs et les éventuelles nouveautés, ainsi qu'un guide méthodologique. Les journalistes disposent donc d'une large palette d'éléments qu'ils peuvent relayer, en citant notamment le site de diffusion officiel du Ministère. Sur ce site, des vidéos didactiques expliquent de manière simple comment sont calculés les indicateurs (MENJ-DEPP, 2020). Toute cette documentation, élaborée avec l'aide des services de la communication du ministère, a semble-t-il permis d'améliorer, d'année en année, la prise en main et la compréhension des IVAL.

8. Mathématiques, Numérique et Sciences Informatiques, Sciences Économiques et Sociales, etc.

Figure 4. Des indicateurs partagés pour des finalités différentes



POUR LA PRESSE, LA TENTATION DES PALMARÈS…

Même si les indicateurs n'ont pas été conçus pour cette finalité, des palmarès continuent de fleurir tous les ans au printemps dans la plupart les médias. Parmi les critères utilisés par les journalistes pour produire ces classements, certains sont très éloignés de la philosophie des IVAL, comme le classement de *L'Internaute* (mars 2020), qui calcule, pour chaque lycée, une moyenne des taux bruts. Ou encore celui du *Figaro*, qui met en avant les lycées ayant les meilleurs taux de réussite et de mentions observés. D'autres journaux adoptent une position médiane, en calculant une note pour chaque lycée, à partir à la fois des taux observés et des valeurs ajoutées *(L'Express, L'Étudiant)*. Enfin, certains se sont mieux approprié les principes pensés par les concepteurs des IVAL et leur intérêt. *Le Parisien – Aujourd'hui en France*[9], notamment, calcule la somme des trois valeurs ajoutées, et n'utilise les taux bruts que pour départager les lycées à égalité.

La presse régionale n'est pas en reste, et publie également de nombreux articles. À partir des mêmes données objectives, des classements extrêmement différents sont ainsi publiés. Certains ont encore trop tendance à mettre en avant les lycées favorisés, ce qui peut entraîner des effets pervers. Ils nuisent en particulier à l'attractivité des lycées les plus défavorisés, quand bien même leur valeur ajoutée est positive, et mettent en avant des lycées favorisés, aux valeurs ajoutées parfois négatives. De plus, les parents d'élèves qui s'intéressent à ces classements, mieux informés, sont également parents des enfants au profil plus favorisé. Un classement établi sur des critères biaisés peut alors conduire certains d'entre eux à éviter des lycées pourtant méritants.

« Si tout n'est donc pas parfait dans ce que publient les médias, il y a néanmoins beaucoup d'éléments permettant de nourrir le débat public. »

Si tout n'est donc pas parfait dans ce que publient les médias, il y a néanmoins beaucoup d'éléments permettant de nourrir le débat public. Les palmarès quantitatifs sont en effet fréquemment accompagnés d'éléments qualitatifs : articles approfondis sur la manière dont sont évalués les lycées, interviews de spécialistes ou encore reportages sur le terrain. Ces derniers ont le plus souvent lieu au sein d'établissements ayant obtenu de bons résultats, et apportent un éclairage sur les méthodes qui fonctionnent.

Au cours des premières années de leur diffusion, la majorité des médias utilisaient uniquement le taux de réussite observé pour classer les lycées. Le travail de pédagogie a porté ses fruits, puisqu'ils sont désormais de plus en plus nombreux à utiliser non seulement les autres indicateurs fournis, mais aussi les valeurs ajoutées. Si des efforts restent à faire, le traitement médiatique des indicateurs de performance des lycées est de plus en plus affiné.

En parallèle du traitement qui en est fait par les médias, les indicateurs sont également utilisés comme outils de pilotage interne. Au sein du ministère, de nombreux acteurs institutionnels s'en emparent, notamment les recteurs, les services académiques et les proviseurs de lycées.

9. Le quotidien propose par ailleurs des cartes (*http://etudiant.aujourdhui.fr/etudiant/carte-palmares-des-lycees-le-parisien.html*) sur lesquelles sont positionnés les lycées, avec des couleurs différentes pour chacune des cinq familles décrites dans la ***figure 5***.

LES PROVISEURS DE LYCÉES, À LA FOIS UTILISATEURS ET OBJETS DE L'ÉVALUATION

À travers les résultats de leur lycée, c'est aussi, d'une certaine manière, l'action des proviseurs qui est évaluée. Il est donc doublement nécessaire de les faire adhérer au dispositif.

Dès début janvier, soit deux mois avant la diffusion officielle, les proviseurs ont accès à un site de « validation », *via* lequel ils peuvent consulter les résultats provisoires de leur établissement. Cette phase, qui dure environ un mois, leur permet de faire part à la DEPP de leurs remarques, voire de contester les chiffres.

La plupart des remarques relèvent d'incompréhensions concernant la méthodologie ou les concepts utilisés. En particulier, la notion d'élève présent au baccalauréat est moins facile à appréhender qu'elle n'y paraît. Un élève est en effet comptabilisé comme présent s'il a au moins une note à son actif relevant du contrôle en cours de formation ou d'une épreuve terminale. Cela permet d'éviter que certains lycées ne fassent artificiellement gonfler leurs résultats en dissuadant leurs élèves les plus faibles de se présenter aux épreuves terminales. Tous les ans, des proviseurs dont un ou plusieurs élèves ont quitté le lycée en cours d'année s'en désolent, notamment dans la voie professionnelle. Ces élèves, s'ils ont obtenu une note de contrôle continu, sont en effet comptabilisés comme présents, et donc en échec.

Selon les années, entre 100 et 200 demandes de révision (sur environ 4 300 lycées) parviennent à la DEPP. Quelques dizaines d'entre elles sont acceptées, par exemple dans le cas d'élèves malades n'ayant pu se présenter aux épreuves (un justificatif médical est demandé). La grande majorité des proviseurs valident ainsi les chiffres calculés, et peuvent préparer leur communication auprès des parents d'élèves et des inspecteurs d'académie. Cette phase d'échanges est par ailleurs essentielle, car les remarques des proviseurs donnent des pistes intéressantes d'évolutions. Celles-ci peuvent concerner la prise en compte des fusions d'établissements, de l'offre de formation, ou encore d'élèves nécessitant un traitement particulier[10].

DES OUTILS PROPOSÉS AUX SERVICES ACADÉMIQUES

Les académies sont les circonscriptions administratives de référence de l'Éducation nationale, il en existe une trentaine, et chacune d'entre elles abrite un Service statistique académique (SSA). Au sein d'un réseau animé par la DEPP, les SSA jouent un rôle important dans l'élaboration d'informations d'aide à la décision et au pilotage. Lors de la diffusion des IVAL, ces services peuvent être amenés à produire des notes, à destination du recteur, sur les résultats des lycées de leur académie. La DEPP joue un rôle d'assistance et répond aux éventuelles sollicitations. Elle propose par exemple différents types de représentations graphiques, parmi lesquelles des nuages de points croisant les valeurs ajoutées : dans l'exemple de la ***figure 5***, la répartition des lycées de l'Académie A est très homogène. En comparaison, l'académie B contient davantage de lycées sélectifs et en deçà des attentes. Dans d'autres académies, ce sont les lycées performants qui sont les mieux représentés. Les seuils qui permettent de définir les catégories[11] ont été choisis de manière à obtenir

10. Cours du soir, dispositifs de lutte contre le décrochage scolaire, etc.
11. Valeurs ajoutées comprises entre −3 et +3 pour les lycées « neutres », par exemple.

une répartition homogène au niveau national. Dans la voie professionnelle, où la dispersion des valeurs ajoutées est plus importante, les seuils sont d'ailleurs plus élevés. Dans une optique d'aide au pilotage, ce type de représentation est plus pertinent qu'un classement unidimensionnel.

ÉLARGIR L'UTILISATION DES IVAL POUR LE PILOTAGE DU SYSTÈME ÉDUCATIF

Au niveau national, les IVAL ont certes pour but d'objectiver le débat public, mais au-delà, quelles sont les suites données à leur publication ? Quelques exemples montrent leur contribution au pilotage du système éducatif.

En 2015, l'Inspection générale de l'Éducation nationale (IGEN) s'est saisie des indicateurs pour tenter d'identifier les spécificités que présentent, dans leur mode de fonctionnement, les lycées à forte valeur ajoutée (IGEN et IGAENR, 2015). Pour ce faire, les auteurs ont réalisé une étude de terrain auprès de 71 établissements aux valeurs ajoutées très positives ou très négatives. Il en est ressorti que la valeur ajoutée dépend toujours d'une conjonction de facteurs, elle-même variable selon les lycées. Parmi les nombreux facteurs susceptibles d'engendrer des valeurs ajoutées positives, on citera par exemple : des équipes fédérées autour d'un projet pédagogique, l'implication des enseignants, un accompagnement personnalisé des élèves, un degré d'exigence affirmé ou encore un climat scolaire apaisé.

> *« Si les IVAL sont bien connus des équipes de direction, ils n'en sont pas pour autant largement diffusés au sein des établissements. »*

Au fil des entretiens avec les proviseurs et les enseignants, les inspecteurs ont par ailleurs constaté que, si les IVAL sont bien connus des équipes de direction, ils n'en sont pas pour autant largement diffusés au sein des établissements. De toute évidence, les lycées dont la valeur ajoutée est positive ont davantage tendance à communiquer sur le sujet, que ce soit en interne, auprès des parents d'élèves ou de la presse régionale. Globalement, les IVAL restent cependant méconnus des professeurs, qui n'y voient que des indicateurs parmi d'autres, sans percevoir la spécificité du calcul « en valeur ajoutée ».

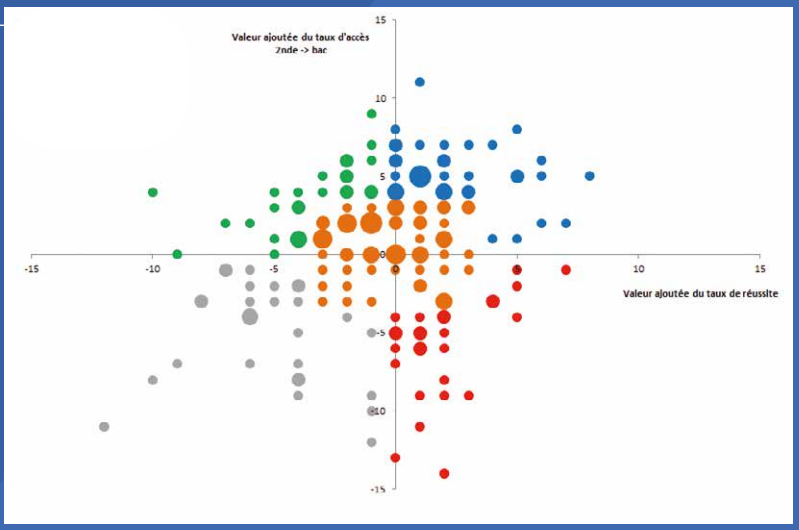
Au-delà des indicateurs de valeur ajoutée, se pose la question de l'évaluation globale des lycées. Les IVAL ne peuvent en effet résumer à eux seuls les qualités pédagogiques d'un établissement. Pointant l'absence de leur évaluation systématique, un nouveau rapport de l'Inspection générale estimait en 2017 qu'il fallait passer de l'évaluation des résultats aux examens à celle du lycée dans sa globalité (IGEN et IGAENR, 2017). Pour traiter d'autres aspects de la vie d'un lycée, comme le climat scolaire par exemple, les auteurs invitent à s'inspirer des indicateurs de valeur ajoutée. La prise en compte des éléments de contexte (caractéristiques des élèves accueillis, environnement socio-culturel, etc.) apparaît ainsi comme étant nécessaire à toute évaluation.

La famille des indicateurs de valeur ajoutée va par ailleurs s'agrandir en fin d'année 2020, avec la diffusion, pour chaque lycée professionnel et centre de formation d'apprentis, d'un taux d'insertion dans l'emploi. Comme pour les IVAL, ces taux seront accompagnés de valeurs ajoutées, et des indicateurs complémentaires seront mis à disposition : taux de poursuite et d'interruption d'études et taux de rupture des contrats d'apprentissage.

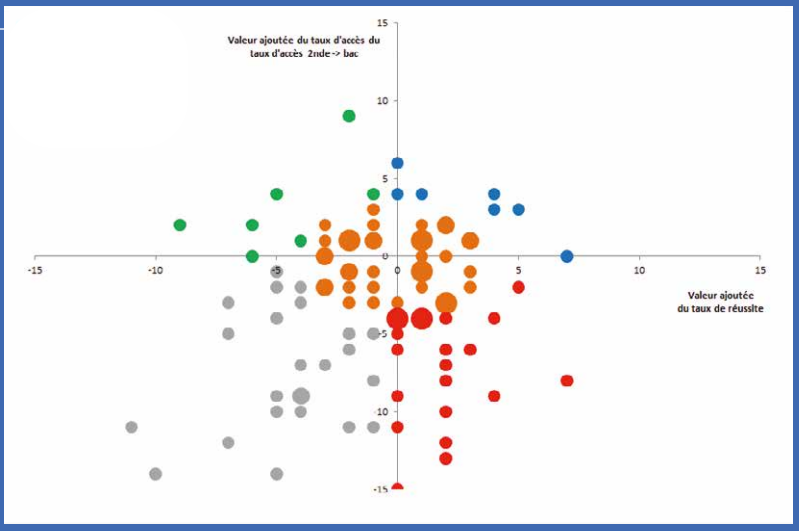
Figure 5. Représenter les performances des lycées pour les comparer et les caractériser

Ces nuages, dont chaque point représente un nombre d'établissements, permettent de distinguer cinq types de lycées :

Académie A



Académie B



- les *neutres*, qui ne contribuent ni plus, ni moins, à la réussite de leurs élèves que la moyenne des lycées leur ressemblant en termes de profils d'élèves accueillis ;
- les *accompagnateurs*, dont les élèves mettent peut-être un peu plus de temps à obtenir le baccalauréat, mais quittent moins souvent le lycée en cours de scolarité ;
- les *sélectifs* en cours de scolarité, que les élèves quittent plus souvent, mais où ils ont une meilleure réussite à l'examen final, pour ceux qui y restent ;
- les *en deçà des attentes*, qui ont de moins bons résultats à la fois en termes de réussite et d'accès au baccalauréat, compte tenu du profil de leurs élèves ;
- les *performants*, que les élèves quittent moins souvent en cours de scolarité et où ceux qui se présentent à l'examen le réussissent plus souvent.

> *« Le prochain défi d'importance à relever, pour la DEPP, sera d'adapter la méthodologie des IVAL aux collèges. »*

En plus d'enrichir l'évaluation des établissements, ces indicateurs permettront d'éclairer les choix des jeunes de l'enseignement professionnel, et contribueront à la réflexion sur l'articulation entre formation et emploi dans chaque territoire. Le prochain défi d'importance à relever, pour la DEPP, sera d'adapter la méthodologie des IVAL aux collèges. Une perspective rendue possible par la mise en place, depuis 2017, d'évaluations exhaustives des élèves de sixième. Cette estimation du niveau des élèves à l'entrée au collège était indispensable. Elle devrait permettre, dans les années à venir, de calculer des indicateurs de valeur ajoutée pour chaque collège de France, étendant ainsi à tout le secondaire une démarche et une méthode qui ont fait leurs preuves.

BIBLIOGRAPHIE

BUISSON-FENET, Hélène, 2019. *Piloter les lycées – Le tournant modernisateur des années 1990 dans l'éducation nationale.* 24 octobre 2019. Éditions PUG, Collection Libres cours Politique. ISBN 978-2-7061-4281-9.

CYTERMANN, Jean-Richard, 2005. La contribution des outils statistiques et d'évaluation à la modernisation de l'Éducation nationale. In : *Politiques et management public.* [en ligne]. Vol. 23, n° 1, 2005, pp. 91-103. [Consulté le 26 novembre 2020]. Disponible à l'adresse : *https://www.persee.fr/doc/pomap_0758-1726_2005_num_23_1_2264.*

DUCLOS, Marc et MURAT, Fabrice, 2014. Comment évaluer la performance des lycées ? Un point sur la méthodologie des IVAL (Indicateurs de valeur ajoutée des lycées). In : *Éducation & formations.* [en ligne]. Novembre 2014. MENESR-DEPP, n° 85, pp. 73-84. [Consulté le 26 novembre 2020]. Disponible à l'adresse : *https://www.epsilon.insee.fr/jspui/bitstream/1/40674/1/depp_educ_form_2014_85.pdf.*

EVAIN, Franck et ÉVRARD, Lætitia, 2017. Une meilleure mesure de la performance des lycées – refonte de la méthodologie des IVAL (session 2015). In : *Éducation & formations.* [en ligne]. Septembre 2017. MEN-DEPP, n° 94, pp. 91-116. [Consulté le 26 novembre 2020]. Disponible à l'adresse : *https://www.education.gouv.fr/sites/default/files/imported_files/document/DEPP-EF94-2017-article-5-meilleure-mesure-performance-lycees-refonte-methodologie-ival-session-2015_819385.pdf.*

FÉLOUZIS, Georges, 2004. Les indicateurs de performances des lycées, une analyse critique. In : *Éducation & formations.* [en ligne]. Décembre 2004. MENESR-DEP, n° 70, pp. 83-95. [Consulté le 26 novembre 2020]. Disponible à l'adresse : *https://www.epsilon.insee.fr/jspui/bitstream/1/40732/3/depp_educ_form_2004_70.pdf.*

FOUGÈRE, Denis, GIVORD, Pauline, MONSO, Olivier et PIRUS, Claudine, 2019. Les camarades influencent-ils la réussite et le parcours des élèves ? Les effets de pairs dans l'enseignement primaire et secondaire. In : *Éducation & formations.* [en ligne]. Décembre 2019. MENJ-DEPP, n° 100, pp. 23-52. [Consulté le 26 novembre 2020]. Disponible à l'adresse : *https://www.education.gouv.fr/sites/default/files/imported_files/document/depp-2019-EF100-article-02_1221886.pdf.*

GIVORD, Pauline et GUILLERM, Marine, 2016. *Méthodologie statistique – Les modèles multiniveaux.* [en ligne]. 21 juillet 2016. Insee, Documents de travail, n° M2016/05. [Consulté le 26 novembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/2022152.*

IGEN et IGAENR, 2015. *Des facteurs de valeur ajoutée des lycées.* [en ligne]. Juillet 2015. Rapport à madame la ministre de l'Éducation nationale, de l'enseignement supérieur et de la recherche, n° 2015-065. [Consulté le 26 novembre 2020]. Disponible à l'adresse : *https://www.education.gouv.fr/sites/default/files/2020-02/2015-065-valeur-ajoute-lycees-510755-pdf-31388.pdf.*

IGEN et IGAENR, 2017. *L'évaluation des établissements par les académies.* [en ligne]. décembre 2017. Rapport à monsieur le ministre de l'Éducation nationale, n° 2017-080. [en ligne]. [Consulté le 26 novembre 2020]. Disponible à l'adresse : *https://www.education.gouv.fr/l-evaluation-des-etablissements-par-les-academies-9410.*

MENJ-DEPP, 2020. Méthodologie des indicateurs de résultats des lycées. In : *site du ministère de l'Éducation nationale et de la jeunesse.* [en ligne]. 21 mars 2020. [Consulté le 26 novembre 2020]. Disponible à l'adresse : *https://www.education.gouv.fr/methodologie-des-indicateurs-de-resultats-des-lycees-11948*.

PONCEAU, Juliette, 2019. *Parcours dans l'enseignement supérieur : du baccalauréat au premier diplôme du premier cycle.* [en ligne]. 21 juin 2019. MESRI DGESIP/DGRI SIES, Note d'information Enseignement supérieur, Recherche & Innovation, n°19-09. [Consulté le 26 novembre 2020]. Disponible à l'adresse : *https://cache.media.enseignementsup-recherche.gouv.fr/file/2019/06/2/NI_19.09_1144062.pdf*.

ROCHER, Thierry, 2016. Construction d'un Indice de position sociale des élèves. In : *Éducation & formations.* [en ligne]. Avril 2016. MENJ-DEPP, n°90, pp. 5-27. [Consulté le 26 novembre 2020]. Disponible à l'adresse : *https://www.epsilon.insee.fr/jspui/bitstream/1/41994/1/depp_educ_form_2016_90.pdf*.

PRISME

DU RÉGIME GÉNÉRAL AU RÉGIME UNIVERSEL, LA MICROSIMULATION COMME OUTIL D'AIDE À LA DÉCISION

Bryan Bellanger et Samuel Goujon***

En France, les nombreux régimes qui composent le système de retraite sont le fruit de sa singularité historique mais aussi les témoins de sa complexité. Ce système connaît régulièrement des évolutions législatives portées par les différentes réformes et décrets, impliquant la mobilisation de maints acteurs de la production statistique. Dans ce contexte, la Cnav – premier régime de retraite français – s'est dotée dès le début des années deux-mille d'un modèle de microsimulation dynamique nommé Prisme.

Contrairement à d'autres modèles, Prisme se distingue par son insertion au sein d'un régime, l'inscrivant dans un usage « métier ». Ainsi il porte un double objectif : de gestion et de prospective. Par ailleurs, Prisme bénéficie de données riches et uniques puisque qu'il s'appuie sur les bases de gestion de la Sécurité sociale.

L'architecture du modèle se décompose en modules construits au plus proche des trajectoires de vie des assurés : démographie, carrière, retraite. Si Prisme est techniquement complexe, pouvant projeter des scénarios à court terme comme à long terme (horizon 2070), sa malléabilité lui permet de s'adapter continuellement aux évolutions du système comme en témoigne l'intégration récente du régime des travailleurs indépendants.

French pension system is based on numerous schemes: this complexity results from its historical singularity. Frequent reforms and decrees generate many changes in the pension system, which has a strong impact on statistical production. In this context, CNAV – entity that manages France's leading pension scheme – developed a dynamic microsimulation model called PRISME in the early 2000s.

Contrary to other models, PRISME is characterized by its insertion within a pension scheme, making it a main tool for operational uses. Thus it's carrying a double objective: management and prospective. Moreover, PRISME benefits from rich and unique sources of data since it is based on the Social Security management databases.

The model consists of different modules built as close as possible to the life paths of the insured: demography, career, retirement. PRISME is technically complex, being able to simulate both short-term and long-term scenarios (horizon 2070). But its is also very flexible, which allows it to adapt continuously to changes in the legal framework, as shown by the recent integration of the self-employed workers scheme in the model.

* Chargé d'études statistiques, Caisse nationale d'assurance vieillesse (Cnav), *bryan.bellanger@cnav.fr*

** Chargé d'études statistiques, Caisse nationale d'assurance vieillesse (Cnav), *samuel.goujon@cnav.fr*

UN SYSTÈME DE RETRAITES MARQUÉ PAR L'HÉRITAGE HISTORIQUE

Le système de retraites français est le fruit d'une volonté postérieure à la deuxième guerre mondiale de construire un système d'assurance vieillesse capable de couvrir l'ensemble de la population. Il s'inscrit dans l'objectif d'offrir aux Français un large système de couverture sociale, permettant de se prémunir de tous les risques : santé, retraites, famille et accidents du travail. À ses débuts, l'existence de catégories auxquelles étaient déjà versées des pensions de retraite[1], doublée des réserves émises par les travailleurs indépendants, a orienté les choix vers un système de retraites par catégories socioprofessionnelles. Il en a résulté une pluralité de régimes, pour la plupart toujours existants, de dimensions bien différentes – allant de quelques centaines de bénéficiaires pour le régime du port autonome de Strasbourg ou celui de la Comédie-Française, à plusieurs millions pour le régime général qui regroupe les salariés du secteur privé. À cela s'ajoutent des réglementations disparates et des règles de calcul des pensions propres à chaque régime.

> *« Ce contexte de vieillissement de la population soulève des questions sur des enjeux majeurs : les financements, les dépenses et donc l'équilibre financier de ces régimes, pensés à une époque où la répartition entre le nombre de cotisants et les nombre de retraités était toute autre. »*

Depuis, les évolutions sociétales et en particulier les améliorations des conditions de travail, des conditions sanitaires combinées aux progrès de la médecine, ainsi que le vieillissement des générations nombreuses du *« baby-boom »* ont amené une part de plus en plus massive d'individus aux âges auxquels ils pouvaient non seulement prétendre à la retraite, mais aussi et surtout en bénéficier plus longtemps. Ce contexte de vieillissement de la population soulève des questions sur des enjeux majeurs : les financements, les dépenses et donc l'équilibre financier de ces régimes, pensés à une époque où la répartition entre le nombre de cotisants et les nombre de retraités était toute autre.

Plusieurs réformes réglementaires ont donc été menées ces 30 dernières années afin d'assurer la pérennité financière des régimes de retraite. En 2000, le **Conseil d'orientation des retraites** (COR) a été créé afin d'étudier, d'analyser et de suivre les perspectives à moyen et long terme de chaque régime, dans le but d'avoir une vision globale de l'ensemble du système de retraites. Ce conseil n'a pas qu'une vision prospective, son rôle est aussi de vérifier que les décisions réglementaires prises auparavant conduisent bien aux effets escomptés plusieurs années voire décennies plus tard.

Cette ambition de disposer d'une vision éclairée du devenir du système de retraites a motivé au sein de chaque régime, notamment auprès de ceux pour lesquels les enjeux financiers sont plus prépondérants, la création de nouveaux outils de projection. Ceux-ci doivent se montrer aussi souples que pratiques, aussi modulables que lisibles et en mesure de s'adapter aux éventuelles modifications à venir.

Parmi toute la palette d'outils envisageables, la microsimulation dynamique, malgré quelques limites, répond bien à ces attentes (Blanchet, 2020). À l'instar d'autres institutions

1. Fonctionnaires et cadres du secteur privé.

sur des sujets voisins, comme l'Insee avec le modèle Destinie (Blanchet *et alii,* 2011), la Drees[2] avec TRAJECTOiRE (Cheloudko et Martin, 2020) ou la Direction générale du Trésor avec Aphrodite, la Caisse nationale d'assurance vieillesse (Cnav), en charge du régime général des salariés, a choisi de se doter d'un modèle de microsimulation dynamique appelé Prisme.

PRISME, UN MODÈLE POUR SE PROJETER À L'HORIZON 2070

Juste après la réforme de 2003[3] ***(encadré 1)***, et après avoir contribué aux chiffrages préalables à celle-ci, la Cnav a décidé de se munir d'un nouveau modèle de projection. Le projet était ambitieux. Le modèle devait servir à l'ensemble des travaux de prospective de la branche vieillesse de la Sécurité sociale : depuis des prévisions à l'année N pour la Commission des comptes de la Sécurité sociale (CCSS, 2020) et des prévisions quadriennales pour les projets de loi de financement de la Sécurité sociale, jusqu'aux projections à long terme pour l'analyse de l'évolution de notre système dans les 50 prochaines années, mais aussi et surtout la simulation des futures réformes afin d'aider les pouvoirs à élaborer ces dernières. Parallèlement, le modèle devait s'insérer dans la logique de gestion d'un organisme dont le cœur de métier est la retraite et être en mesure de représenter un appui à la détermination et la compréhension des charges prévisionnelles d'activité.

> *« Le modèle devait s'insérer dans la logique de gestion d'un organisme dont le cœur de métier est la retraite et être en mesure de représenter un appui. »*

Il aura fallu un an et demi aux experts de la Cnav pour concevoir et développer une première version de ce modèle. Il est baptisé **Prisme**, pour « **Projection des Retraites Individuelle, Simulations, Modélisation et Évaluations** ». Au printemps 2005, Prisme fournit ses premières projections à l'horizon 2050 pour alimenter le rapport du Conseil d'orientation des retraites sur l'équilibre financier du système de retraite français. Depuis, Prisme continue chaque année d'alimenter les rapports du COR avec des projections étendues à l'horizon 2070 (COR, 2020).

LE CHOIX DE LA MICROSIMULATION DYNAMIQUE

Prisme est défini comme un modèle de microsimulation dynamique, par opposition aux modèles dits statiques. Les modèles statiques illustrent une situation ou un phénomène étudié à un instant t, de manière immuable et figée dans le temps. Or, quand il s'agit de réforme des retraites, les mesures mises en place s'évaluent bien souvent sur une période de plusieurs années, voire décennies, ou générations.

La microsimulation dynamique permet quant à elle d'introduire la notion de temps et ce qui peut en découler, comme les évolutions législatives, les mutations sociétales (allongement des études par exemple, modification du marché du travail), les dynamiques démographiques (hausse de l'espérance de vie, baisse de la fécondité), etc.

2. Direction de la recherche, des études, de l'évaluation et des statistiques, service statistique du ministère des Solidarités et de la santé.
3. Réforme mise en place au 1er janvier 2004 suite à l'adoption de la loi du 21 août 2003 (voir les références juridiques en fin d'article).

L'idée fondamentale de la microsimulation consiste à modéliser et simuler les évènements et leur probabilité de survenance au niveau de l'individu. Dans le cas de Prisme, l'unité de projection du modèle est en l'occurrence l'assuré. La plupart des modèles de microsimulation se basent sur un échantillon d'individus, même si l'exhaustivité n'est pas ou n'est plus un frein, compte tenu des moyens informatiques actuels. Prisme repose sur un échantillon au 1/20e extrait des bases de gestion de la Cnav : celles-ci comprennent toutes les personnes nées en France hexagonale et en Outre-mer, ainsi que les personnes nées à l'étranger et ayant été immatriculées en France à un moment de leur vie, qu'elles soient encore ou non présentes sur le territoire français. Ces bases de gestion exhaustives, puisque découlant de l'activité même du métier retraite de la Cnav, sont d'une richesse infinie tant elles offrent une diversité de profils qui représentent autant de matière pour alimenter le modèle. L'échantillon compte 5,5 millions d'individus.

> *« Prisme repose sur un échantillon au 1/20e extrait des bases de gestion de la Cnav. »*

Au fil du temps et notamment avec l'histoire sociale récente, le modèle de la Cnav a pu démontrer toute sa pertinence et il s'est de plus forgé, grâce à sa polyvalence et sa malléabilité, une véritable identité dans les évaluations des différentes réformes de retraites pour devenir un outil essentiel d'aide à la décision des politiques publiques dans ce domaine.

PRISME : UN MODÈLE EN CONSTANTE ÉVOLUTION

Depuis 2005, Prisme est l'outil central des travaux de prospective menés par la Cnav sur les retraites de base du régime général. Même si les méthodes et l'architecture globale du modèle sont demeurées les mêmes, Prisme n'a cessé d'évoluer tout au long de ces quinze dernières années : des évolutions évidemment en lien avec les changements de réglementations, pour actualiser les informations sur les assurés, améliorer la modélisation des différents évènements, mais également des évolutions du champ couvert ; car progressivement, Prisme est devenu un modèle couvrant l'intégralité du système de retraite en France.

L'utilisation de l'outil pour la sphère « métier » de la Cnav soumet ses concepteurs à l'obligation d'une veille constante sur la législation concernant les retraites, afin que les modifications même les plus récentes soient intégrées.

L'axe majeur des améliorations du modèle reste néanmoins celles liées aux fortes évolutions réglementaires instaurées par les réformes successives ***(encadré 1)***. Chaque réforme apportant de nouvelles règles et des mesures spécifiques, leur implémentation dans Prisme nécessite leur retranscription en langage informatique. Cette étape est d'autant plus délicate que les normes nouvelles ne concernent pas obligatoirement tous les individus et que certains dispositifs sont soumis à des critères d'éligibilité très fins, éléments que le modèle doit alors également projeter sur plusieurs décennies. La mise en place du dispositif de retraites anticipées pour carrière longue lors de la réforme de 2003 en est un bon exemple, en fixant des modalités différentes selon l'âge de départ, l'âge de début d'activité, la durée cotisée ou encore la durée validée, etc. Si la réforme de 2003 a été intégrée dès la conception de Prisme, les suivantes (2010, 2012 et 2014) ont permis de confirmer la souplesse et la flexibilité du modèle dans la prise en compte de nouveaux paramètres.

Encadré 1. Les principales réformes des retraites

Loi n° 2003-775 21 août 2003

- Hausse de la durée d'assurance en fonction de l'augmentation de l'espérance de vie à 60 ans et alignement de la durée de référence prise en compte pour le calcul des pensions sur la durée nécessaire pour bénéficier du taux plein
- À partir de 2009, hausse de la durée d'assurance d'un trimestre par an pour atteindre 164 trimestres en 2012
- Mise en place d'un dispositif pour les carrières longues permettant aux individus ayant commencé à travailler jeune de partir en retraite anticipée (avant l'âge légal)
- Instauration d'un dispositif de surcote et allégement de la décote

Loi n° 2010-1330 9 novembre 2010

- Relèvement progressif de l'âge légal : de 60 ans à 62 ans (à raison de 4 mois pour les individus nés entre le 1/07/1951 et le 31/12/1951, puis 5 mois supplémentaires jusqu'à la génération 1955)
- Relèvement progressif de l'âge au départ sans décote : de 65 à 67 ans (au même rythme que la hausse de l'âge légal)
- Mise en place d'un dispositif de retraite anticipé pour pénibilité

Décret n° 2012-847 2 juillet 2012

- Élargissement du dispositif des retraites anticipés pour carrière longue aux individus justifiant d'un début d'activité avant 20 ans (départ à 60 ans).

Loi n° 2014-40 20 janvier 2014

- Hausse de la durée d'assurance à raison d'un trimestre tous les 3 ans pour atteindre 172 trimestres pour la génération 1973
- Création du compte personnel de prévention de la pénibilité
- Revalorisation des pensions au 1er octobre (au lieu du 1er avril)
- Acquisition d'un trimestre d'assurance retraite après avoir cotisé sur un revenu équivalent à 150 h de Smic (contre 200 h auparavant)

(voir les références juridiques en fin d'article)

> *C'est d'ailleurs cette nécessité de pouvoir faire évoluer le modèle et d'être au plus proche de la réglementation qui a conduit en partie au choix de la microsimulation.*

C'est d'ailleurs cette nécessité de pouvoir faire évoluer le modèle et d'être au plus proche de la réglementation qui a conduit en partie au choix de la microsimulation. En particulier, dans le mode de calcul de la pension des régimes en annuités, il n'y a ni linéarité, ni proportionnalité rendant difficile voire impossible l'implémentation précise de la législation dans un modèle macro. Pour illustrer ce principe, le passage de 200 à 150 heures de SMIC nécessaires pour cotiser un trimestre n'aura pas le même effet pour tout le monde :

- une personne ayant déjà cotisé ses quatre trimestres annuels sur la base de 200 heures ne sera aucunement concernée par cette évolution pour sa retraite de base ;
- en revanche, une personne qui cotisait avant la réforme seulement trois trimestres sur la base des 200 heures de SMIC devrait dorénavant en cotiser quatre ;
- et celle qui avait seulement deux trimestres cotisés pourrait parvenir à trois voire quatre trimestres selon ses revenus.

Avec la microsimulation, la possibilité de suivre les parcours de vie au cas par cas se révèle être bien plus précise que l'application d'un correctif à un effectif agrégé.

La microsimulation constitue par ailleurs un atout pour effectuer des analyses fines basées sur des données individuelles réelles. Cette méthode permet la construction d'indicateurs avec tous les niveaux d'analyse et d'agrégation de population souhaités, par exemple : pensions moyennes, masses financières, analyses par cycle de vie, étude de certains éléments de pension ou de carrière, analyse des gagnants/perdants en cas d'estimation des effets d'une réforme, etc. Cela prend tout son sens lorsque l'instauration d'une mesure vise par exemple à améliorer les conditions d'une population particulière, comme ce fut le cas pour les bénéficiaires du dispositif de retraite anticipée pour carrière longue ou encore la mise en place du compte pénibilité, et à en étudier le suivi dans le temps. Ici aussi, être en mesure d'affiner ces indicateurs est une grande aide pour la fonction métier de la Cnav.

La microsimulation ouvre donc des possibilités d'analyses innombrables, mais ces analyses n'ont de valeur que si les données sont riches (y compris sur le passé – pour mieux estimer le futur) et le nombre d'individus dans l'échantillon important. C'est bien sur ces aspects que Prisme a bâti sa robustesse. Et ceci a été possible, car Prisme est développé par un régime de retraite qui s'appuie sur ses données de gestion et utilise aussi le modèle pour sa gestion.

LE RÉGIME GÉNÉRAL : UN CHAMP D'APPLICATION DÉJÀ TRÈS LARGE

Initialement, le modèle a été développé sur le champ des salariés du privé, affiliés au régime général. Ce champ peut paraître relativement restreint au regard du nombre de régimes de retraites en France (42 régimes[4]). Néanmoins, il s'agit du principal régime de retraite français, représentant en 2018 plus de 90 % des affiliés à un régime de retraite, et près de 40 % du total des prestations vieillesses versées[5].

4. Voir à ce sujet l'article sur le modèle TRAJECTOiRE (Cheloudko et Martin, 2020).
5. Il suffit d'avoir cotisé un seul trimestre en tant que salarié du secteur privé pour ouvrir des droits au régime général et prétendre ensuite à une pension de retraite.

> *« Prisme doit en effet simuler tous les éléments qui peuvent avoir une incidence sur le montant futur de la pension d'un individu. »*

Qui plus est, pour projeter les futurs cotisants et retraités, il est indispensable de retenir une population plus large que les seuls cotisants actuels de ce régime. Cette population, la Cnav la connaît puisqu'elle gère pour la Cnam (Caisse nationale d'assurance maladie), l'ensemble des immatriculés en France (qu'ils aient ou non un droit retraite), *i.e.* toute personne qui a un numéro de Sécurité sociale (voir *infra*).

Au-delà de la simple population à modéliser, Prisme doit en effet simuler tous les éléments qui peuvent avoir une incidence sur le montant futur de la pension d'un individu. Par exemple, dans l'optique de parvenir à obtenir un nombre de trimestres cotisés ou validés tous régimes confondus, il est primordial pour les *poly-affiliés* de déterminer la partie de la carrière ne relevant pas du régime général. C'est pour cette raison que le modèle s'appuie sur toutes les données possibles, qu'elles soient propres à la Cnav ou bien externes.

DES DONNÉES INDIVIDUALISÉES EXPLOITABLES, RICHES ET VARIÉES

La Cnav dispose de fichiers de gestion qui regroupent les éléments de carrière des salariés au régime général et permettent le paiement des retraites (DSPR, 2020 ; COR, 2020). Ces bases contiennent l'intégralité des personnes qui sont ou ont été affiliées à la Sécurité sociale, soit plus de 109 millions d'individus vivants et décédés (mais pouvant encore ouvrir des droits à réversion)[6]. Deux référentiels sont mobilisés :

- le Système national de gestion des identités (SNGI), pour les données démographiques,
- et le Système national de gestion des carrières (SNGC+[7]), pour les éléments de carrière des assurés.

Ce sont des bases de données individuelles, et c'est justement là que réside toute la finesse et la richesse de leur contenu. La collecte des données individuelles débute dès le certificat de naissance (ou au moment de l'immatriculation pour les personnes nées hors de France). Au cours de la carrière professionnelle, elle se poursuit avec un fin niveau de détail sur les revenus, les employeurs, les caractéristiques de l'emploi occupé, le type de contrat, etc., afin de garantir les droits des assurés à hauteur de leurs contributions. De manière plus générale, on distinguera trois grandes temporalités dans l'ensemble des bases de données mobilisées : la carrière, la liquidation et la prestation.

Le SNGI[8] répertorie les états civils des assurés nés en France et des personnes nées à l'étranger qui relèvent d'un régime français de Sécurité sociale. Pour les personnes nées en France, il est alimenté quotidiennement par le RNIPP (Répertoire national d'identification des personnes physiques) géré par l'Insee. À l'inverse, le SNGI alimente le RNIPP pour les personnes nées hors de France. C'est à partir du SNGI que sont tirés au sort les individus qui seront présents dans la table initiale du modèle ***(figure 1)***.

6. Ces effectifs sont supérieurs à la population totale française : ils regroupent toute personne ayant cotisé au moins un trimestre, née ou non en France et présente ou non sur le territoire français, décédée ou non.
7. SNGC+ car en 2017 avec la mise en place de la liquidation unique pour les régimes alignés (LURA) les informations de revenus des indépendants et salariés agricoles ont été intégrées au SNGC, venant compléter les informations du régime général.
8. Créé par le décret n° 2018-390 (voir les références juridiques en fin d'article).

Les données concernant l'activité professionnelle des salariés[9] proviennent du SNGC+. Il mémorise la totalité de la carrière des assurés sociaux pour le calcul de leur retraite au régime général : salaires cotisés, trimestres validés au régime général et dans les autres régimes, périodes de chômage, de maladie, etc. L'information sur les salaires provient des déclarations sociales nominatives[10] mais aussi de toutes les déclarations qui peuvent être effectuées par les employeurs dans le cadre de leurs obligations légales. Les données relatives au chômage sont fournies par Pôle emploi, les données maladies sont transmises par la Cnam et les cotisations AVPF (Assurance vieillesse des personnes au foyer) par la Cnaf (Caisse nationale des allocations familiales).

À ces données issues des référentiels nationaux, s'ajoutent celles provenant du Système national des statistiques des prestataires (SNSP), base de données statistiques exhaustive qui contient l'ensemble des retraités du régime général.

Par ailleurs, d'autres données externes sont mobilisées pour compléter et enrichir ponctuellement celles de la Cnav. Car même si ces dernières sont riches, tant en termes de variables qu'en termes d'individus, elles sont circonscrites aux besoins métier du régime général. Or, les nouveaux besoins de simulation requièrent d'intégrer toujours plus d'informations, que ce soit pour gagner en précision ou pour suivre le flux réglementaire. Des appariements sont alors mis en œuvre, avec des méthodes spécifiques (pour plus de détail, voir ***encadré 2***).

PRISME SIMULE LA «VRAIE VIE»

> *« Le principal objectif est de prédire la date de liquidation et le niveau de pension des assurés à un horizon défini. »*

La première fonction de Prisme est de permettre la simulation de trajectoires individuelles : des périodes d'emploi, des transitions de carrière, des salaires et cotisations, l'arrivée d'un enfant, les accidents de vie (invalidité, accident de travail, décès) en somme ce que l'on appelle des événements. Le principal objectif est de prédire la date de liquidation et le niveau de pension des assurés à un horizon défini, afin d'obtenir certains indicateurs tels que les masses de pensions versées et les recettes que perçoit le régime de retraite.

Pour effectuer une projection avec Prisme, il est nécessaire d'utiliser des informations « sur le passé » qui servent à la fois :

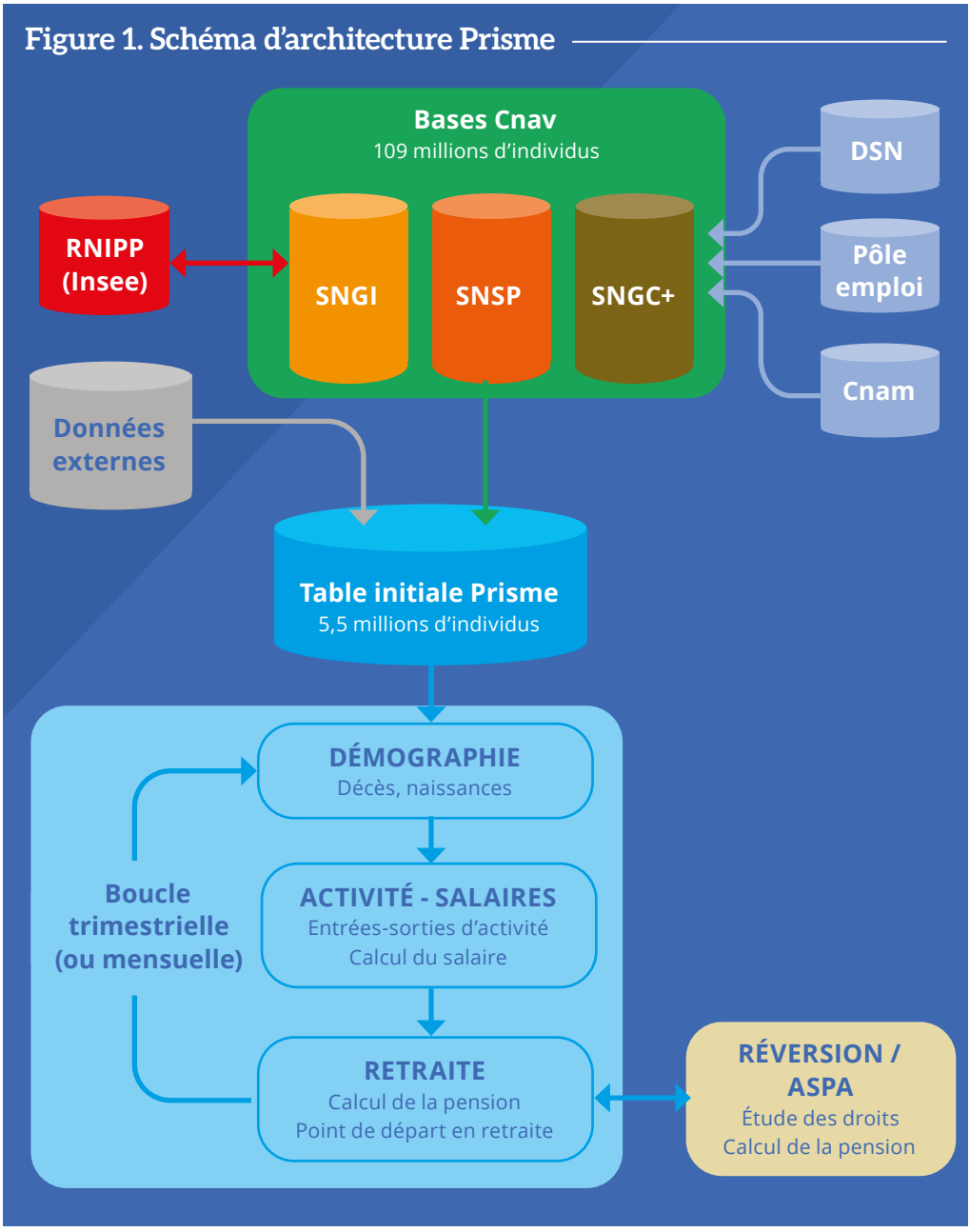
- à compléter le profil actuel des individus : quels sont les éléments de leur vie qui peuvent avoir une incidence future sur la constitution de leurs droits, sur l'âge estimé auquel ils pourraient partir en retraite, etc. ;
- à estimer les événements dans le futur, et ainsi assurer une cohérence et une relative continuité des logiques et structures socio-professionnelles des assurés comme des prestataires des régimes de retraite.

9. Les reports au compte des individus de l'ensemble des régimes sont présents dans la base, ainsi que les salaires et revenus des régimes alignés. Cependant, les salaires et revenus des autres régimes n'y figurent pas.
10. Voir les articles sur la DSN parus dans le numéro N1 du Courrier des statistiques, et notamment (Humbert-Bottin, 2018).

LA TRIMESTRIALISATION : UN GAGE DE PRÉCISION

Le moment auquel intervient un événement est capté dans Prisme, à la fois dans la carrière de l'individu, mais également en infra-annuel : c'est ce qu'on appelle la *trimestrialisation.* Car Prisme fonctionne sur un pas trimestriel, sauf pour les départs en retraite et les décès

Figure 1. Schéma d'architecture Prisme



> *Atout fort du modèle, la trimestrialisation place les événements au plus près de la législation actuelle des régimes de retraite.*

qui sont estimés mensuellement, alors que la plupart des modèles de projection existants dans le système de retraite français fonctionnent sur un pas annuel[11].

Atout fort du modèle, la trimestrialisation place les événements au plus près de la législation actuelle des régimes de retraite (particulièrement pour le régime général) qui repose en effet sur une unité : le trimestre. Ainsi chaque événement de carrière, du passé, comme du futur est estimé individuellement trimestre par trimestre. En ce sens, trimestrialiser permet de reconstituer ou de simuler une plus grande complexité des carrières en infra-annuel. La situation d'activité des individus est ainsi plus représentative de la réalité des situations de vie des assurés : des périodes d'activités salariés, des périodes de chômage, des périodes de maladie, etc. le tout au sein d'une même année.

LES TRAJECTOIRES DE VIE POUR STRUCTURER LE MODÈLE

L'architecture de Prisme se veut calquée sur celle de la « vraie vie » ***(figure 1)***. Le programme est composé de trois modules représentant des séquences de vie (modules « Activités et salaires », « Retraite ») ou des informations individuelles démographiques (module « Démographie »).

Les nombreux événements qui composent la vie des individus sont modélisés dans ces modules, et attribués selon la probabilité que chacun a de vivre ou non ceux-ci (naissance, chômage, maladie, décès, etc.). Est constituée ainsi une trajectoire de vie singulière sur laquelle le modèle va s'appuyer pour évaluer – à terme – un moment de départ à la retraite et un montant de pension.

Le module « Démographie »

Sans cesse actualisé, il détermine – entre autres – tout au long de la vie de l'assuré des événements tels que : les naissances, les décès, les migrations (entrées exclusivement) mais aussi la descendance des assurés (nombre d'enfants).

Les simulations de décès sont différenciées selon que l'on soit *prestataire* c'est-à-dire percevant une pension de retraite (quotients de mortalité estimés sur les observations Cnav selon que le prestataire perçoit ou pas une pension d'ex-invalide ou au titre de l'inaptitude) ou *non-prestataire* (projections Insee).

Les naissances pour les hommes sont distribuées de manière aléatoire, tout en respectant un indice conjoncturel de fécondité. Pour les femmes est ajouté un calendrier des naissances, selon une équation logistique tenant compte de facteurs tels que l'âge et le temps écoulé depuis la sortie des études.

Ce module permet aussi d'ajouter les nouveaux assurés nés au cours des années simulées, selon les projections de population de l'Insee.

11. C'est le cas du modèle de microsimulation de l'Agirc-Arrco, ou encore de TRAJECTOiRE, le modèle de la Drees (Cheloudko, Martin, 2020).

> *Douze types de reports sont modélisés, parmi lesquels : les périodes de chômage, de maladie, d'invalidité, de maternité ainsi que les périodes d'activité selon le régime de retraite de base concerné.*

Il permet de définir, pour les actifs, une situation par rapport à l'emploi à chaque trimestre, que l'on appelle un *report*. Il caractérise ainsi la situation de l'assuré mais aussi sa date d'entrée sur le marché du travail.

Douze types de reports sont modélisés, parmi lesquels : les périodes de chômage, de maladie, d'invalidité, de maternité ainsi que les périodes d'activité selon le régime de retraite de base concerné[12]. Tout au long de la projection, de nouveaux reports sont attribués aux individus selon un pas trimestriel et un enchaînement d'équations ; pour cela le modèle utilise des logits

Encadré 2. L'Appariement optimal: méthode de complétion

Les bases de gestion de la Cnav sont riches et de mieux en mieux alimentées par les autres régimes. Néanmoins, bien que l'amélioration du transfert des données autres régimes soit plus que notable, une partie des données de carrières du Service des retraites de l'État (SRE) sont incomplètes. Ces éléments lacunaires doivent être remplacés par des informations permettant le travail de projection dans Prisme ; on appelle cette étape : la complétion. L'appariement optimal (Lesnard et de Saint Pol, 2006) est la méthode utilisée par la Cnav pour cette phase préliminaire de « préparation des données ».

L'appariement optimal est une technique consistant à mesurer la distance entre des séquences*, deux à deux, et à transformer l'une en l'autre au moyen d'opérations élémentaires. Trois opérations sont possibles** : la substitution d'un élément par un autre, l'insertion d'un élément ou la suppression d'un élément dans la séquence. Un coût*** est affecté à chacune des opérations élémentaires. La distance entre deux séquences est la somme de ces coûts ; l'appariement nécessite que ces derniers soient les plus faibles possibles.

Les données de l'EIC (Échantillon interrégimes de cotisants) de la Drees sont une source d'information précieuse pour la complétion. Ainsi, chaque individu de l'échantillon Cnav présentant une carrière incomplète est comparé aux individus de l'EIC dans l'objectif de l'apparier à un « jumeau » (individu présentant le moins de dissemblances dans la comparaison du parcours professionnel : *i.e.* distance minimale). Ce « jumeau » de carrière permet la substitution des éléments manquants par ceux connus dans le fichier de l'EIC. Ainsi complétée, la base de gestion de la Cnav est exploitable par Prisme.

* Une séquence est un élément longitudinal composé d'un ensemble d'états successifs caractérisant, dans notre cas, le parcours professionnel d'un individu.

** Pour Prisme, sont utilisées les opérations de substitutions.

*** Dans Prisme, un coût de substitution vaut 1/probabilité de la transition. Le coût d'un état vers lui-même est nul.

12. Des conditions de travail pénibles, pour les assurés en emploi salarié, peuvent être simulées afin d'acquérir des points sur leur compte professionnel de prévention et ainsi des droits supplémentaires à la retraite.

multinomiales[13] différenciées par report précédent. En effet les transitions entre les différents états d'activité sont dépendantes de l'état qui précède (Berteau-Rapin, Beurnier et Denayrolles, 2015). Le choix de ce type de modélisation est motivé par la nécessité de contrôler à chaque pas de projection les effectifs se répartissant entre les différents types de reports de carrière. Les probabilités servant à modéliser les transitions peuvent être représentées comme dans la ***figure 2***. Chacune des probabilités est déclinée selon le genre, le lieu de naissance, et l'âge de l'assuré. Elles sont la traduction du « coût de transition » que représente le passage d'un état à un autre pour un assuré ; certaines transitions sont plus fréquentes (moins « coûteuses ») et d'autres beaucoup plus rares, néanmoins leur inclusion est indispensable pour la bonne modélisation de la population en projection.

Afin d'estimer ces logits multinomiales, une période de quelques années est sélectionnée dans le passé, sur laquelle sont observées et mesurées les différentes transitions de carrière. Techniquement, à chaque état sont calculées des probabilités cumulées de transition ; puis un aléa est tiré pour chaque individu. La comparaison de cet aléa aux probabilités cumulées de transition permet d'attribuer un nouvel état, et ainsi de suite.

Toutefois, certaines transitions sont considérées comme impossibles par le modèle[14] ; par exemple, la probabilité pour qu'un assuré percevant à un trimestre donné une pension une allocation *chômage* perçoive au trimestre suivant une allocation *maladie*, est égale à 0.

Ce module permet également de simuler des salaires et revenus, et par ce fait des assiettes de cotisations[15]. Ces éléments sont essentiels pour approcher un montant de pension au niveau micro et des masses de droits propres au niveau macro.

Le module « Retraite »

Il permet d'estimer la probabilité de départ de chacun des assurés – mois après mois – dès leur 55e anniversaire (dans la limite de leur 75 ans). Avec la date de départ ainsi simulée, on peut également calculer des montants de pension au regard des informations de vie, récupérées sur le passé ou modélisées au cours de la projection.

Au total, ce module comptabilise 48 équations logistiques différenciées par âge, sexe, types de départ et dernier régime d'affiliation. Différencier les équations permet d'adapter les variables explicatives (délai depuis la possibilité d'un départ anticipé, distance par rapport au taux plein, situations vis-à-vis de l'emploi, etc.) et ainsi, de saisir de manière plus fine les particularités des profils d'assurés liquidant : décoteurs, surcoteurs, partant à l'âge d'annulation de la décote, etc.

Tout au long de la projection, tant que l'assuré n'a pas liquidé ses droits à la retraite, il rentre dans les différentes équations logistiques selon son profil chaque mois, et si la probabilité estimée à partir du vecteur de variables explicatives est inférieure à un aléa tiré de manière uniforme, l'assuré part à la retraite. À noter qu'une reprise d'activité salariée au régime général, après le passage à la retraite, constitue également un événement modélisé dans le cadre du cumul emploi-retraite.

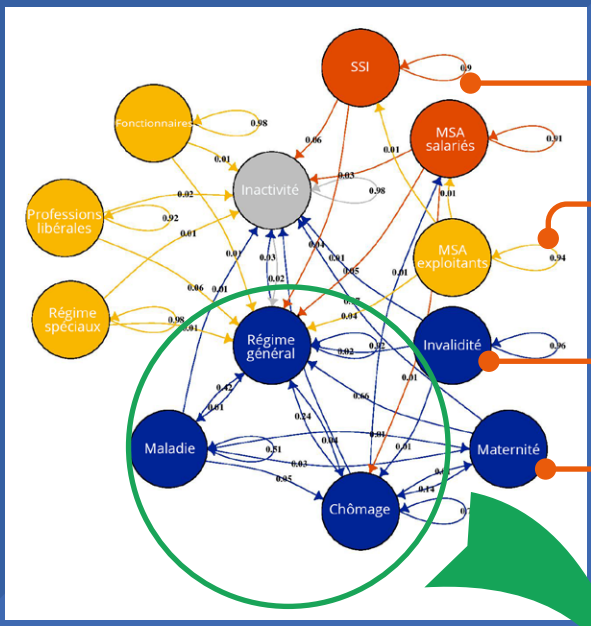
13. La régression logistique multinomiale permet d'estimer la probabilité de survenue d'un événement pour une variable avec plus de deux modalités compte tenu des caractéristiques individuelles.
14. Transitions impossibles d'un point de vue légal ou trop rares pour être estimées avec fiabilité.
15. Une assiette de cotisations est l'ensemble des éléments de base servant au calcul des cotisations.

LA PRÉOCCUPATION DE FINESSE ET DE CONTRÔLE DU MODÈLE…

Au-delà de la colonne vertébrale de Prisme, démographie – carrière – retraite, la recherche de finesse des modélisations et le contrôle même du modèle de microsimulation constituent un travail important.

La prise en compte de certaines prestations qui ne concernent qu'une partie de la population, telles que la réversion ou l'Allocation de solidarité aux personnes âgées (ASPA ou minimum vieillesse), témoigne de cette intention. En effet, bien que ces masses de

Figure 2. Schéma des probabilités de passage d'un état à un autre dans Prisme

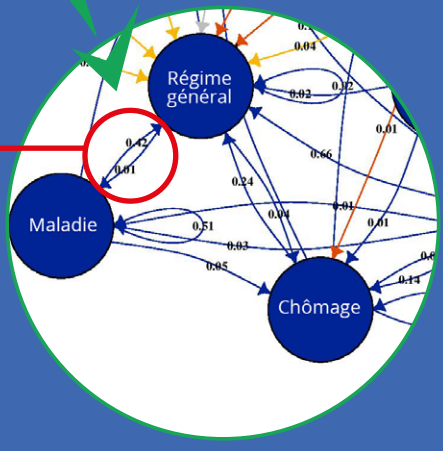


Chaque flèche du diagramme représente une **transition** possible d'un état à l'autre

Le chiffre derrière chaque flèche représente la **probabilité** « de passage » vers le prochain état

Les cercles représentent les différents **états** au titre desquels un individu peut être « assimilé » (un régime de retraite, une situation professionnelle, etc.)

Exemple :
après un trimestre validé au titre d'un salaire au RÉGIME GÉNÉRAL, la probabilité de « reporter » le trimestre suivant au titre d'une période assimilée MALADIE est de 1 % (0,01). La transition inverse est de probabilité 42 % (0,42).



prestations représentent plus de 10 % des prestations versées par la Cnav, Prisme est un des rares modèles français à en estimer les montants (Di Porto et Ghernaout, 2020). À l'instar des droits directs, le modèle a récemment été élargi à l'ensemble des régimes de retraite français dans le cadre des simulations de réforme systémique. Ainsi, lorsque le décès d'un conjoint survient au sein d'un couple marié, une pension dite *de réversion* peut être attribuée au conjoint survivant sous certaines conditions (âge, ressources…). Celle-ci correspond à une partie de la pension du conjoint décédé (54 % au régime général). Ces pensions sont simulées dans Prisme par estimation des mariages, du veuvage, de la pension du conjoint décédé et des ressources du conjoint survivant. Il en va de même, pour l'ASPA, allocation attribuée aux retraités aux faibles ressources. Prisme intègre son calcul pour les populations répondant aux critères d'éligibilité (âge, ressources, résidence, situation familiale, subsidiarité). Outre l'estimation de l'ensemble des pensions de retraite afin d'estimer l'essentiel des ressources, ceci implique de simuler la résidence sur le territoire français des prestataires d'une pension de droit propre, ainsi que le « non-recours » qui s'avère encore élevé pour cette allocation.

Cette recherche de finesse se retrouve aussi dans les nombreuses opérations de calage effectuées afin de respecter les hypothèses macroéconomiques fournies par le Conseil d'orientation des retraites ou la Direction de la Sécurité sociale (DSS) : évolution du nombre de cotisants par exemple, ou part des différents types de carrière dans la population générale (*poly-pensionnés*, etc.). Ces opérations de calage sont indispensables tant l'outil de microsimulation est sensible : il faut s'assurer continuellement de la cohérence des simulations de long et de court terme, en mobilisant des méthodes de validation et un contrôle accru du modèle.

> *« Ces opérations de calage sont indispensables tant l'outil de microsimulation est sensible. »*

… CAR PRISME N'EST PAS UNE BOULE DE CRISTAL

Compte tenu de sa puissance, un modèle de microsimulation dynamique pourrait avoir tendance à modéliser – dans un effet boule de neige – sa « propre logique ». Afin de contenir le modèle et d'évaluer les résultats obtenus, plusieurs étapes de contrôle ponctuent le travail de projection (Brossard *et alii,* 2016). Elles s'appuient à la fois sur l'évaluation de la cohérence des sorties vis-à-vis des observations de la réalité, comme lors des étapes de « calages », mais aussi par le contrôle des hypothèses macroéconomiques, comportementales et méthodologiques. De nombreux tests de sensibilité des résultats à ces différentes hypothèses sont opérés par les statisticiens lors du traitement. Par ailleurs, en amont même du lancement de Prisme, la validité du modèle est vérifiée lors de l'étape de modélisation des différentes équations logistiques que ce soit par exemple pour les transitions de carrière ou les départs en retraite. La qualité des estimations est notamment vérifiée en comparant de manière rétroprojetée les estimations et les données constatées.

Pour mesurer la sensibilité des résultats aux hypothèses macroéconomiques, de nombreuses variantes sont utilisées. Concrètement, cela revient souvent à modifier les valeurs cibles du taux de chômage ou de l'évolution du salaire annuel moyen par tête[16]. De multiples scénarios

16. Le salaire moyen par tête (SMPT) rapporte les masses salariales brutes versées par l'ensemble des employeurs au nombre de salariés en personnes physiques.

sont analysés au regard de différents indicateurs, tels que les masses de prestations et de recettes, ou les valeurs moyennes des décalages de départ et des variations de pensions.

Les projections de retraite, en particulier lorsqu'il s'agit de simuler une réforme, nécessitent également des tests de sensibilité des résultats aux hypothèses méthodologiques de comportement de départ et de prolongation d'activité. Pour de nombreux modèles statistiques, la problématique d'entrée repose principalement sur le choix d'une approche déterministe ou probabiliste.

L'approche probabiliste entraîne une modification des probabilités individuelles sous-jacentes à l'évolution des paramètres définis par la réforme au cours de la projection. En effet, les événements simulés par le programme *via* de nombreuses modélisations logistiques (voir *supra*) et leurs probabilités d'occurrence dépendent des changements législatifs à travers les variables explicatives des équations logistiques. L'avantage de cette approche est d'offrir de meilleures variabilités individuelles en sortie et d'intégrer des scénarios plus complexes souvent plus proches des situations observées sur le réel.

Cependant, si l'approche probabiliste est celle nativement modélisée dans Prisme, les statisticiens de la Cnav ont aussi recours à **l'approche déterministe**. Cette approche est exclusivement utilisée lors de la simulation de réforme pour l'événement *« liquidation retraite »*, tous les autres événements étant traités en probabiliste. L'approche déterministe présente l'avantage de conditionner la simulation par des hypothèses comportementales préalablement définies. Ainsi, les comportements de départ à la retraite par type d'assuré sont connus *a priori* et la simulation se conforme à ces derniers sans les faire évoluer au fil de la projection, rendant plus aisée la comparaison des situations individuelles et permettant donc la création de scénarios contrefactuels. On retrouve cette approche dans certains modèles de projection tel que le modèle Ines de l'Insee, de la Drees et de la Cnaf pour la simulation des politiques sociales et fiscales (Fredon et Sicsic, 2020).

L'ÉLARGISSEMENT PROGRESSIF AUX AUTRES RÉGIMES DE RETRAITES

La prise en compte des autres régimes dans le modèle ne constitue pas en soi une révolution totale.

Pour pouvoir calculer finement les conditions de ressources au sein des modules *réversion* et *minimum vieillesse* du régime général, l'estimation des pensions des régimes complémentaires Agirc-Arrco et Ircantec était déjà réalisée de manière précise depuis plusieurs années. Et comme déjà mentionné plus haut, depuis ses débuts, le modèle a entrepris de simuler des fragments, voire l'intégralité des carrières cotisées ou validées dans d'autres régimes de base que le régime général. Mais ces informations étaient agglomérées sans aucune distinction : elles sont dorénavant décomposées régime par régime.

Cet élargissement s'est déroulé en deux grandes étapes.

La première fut **la mise en place du dispositif de la LURA**[17], entré en vigueur à partir du 1er juillet 2017 pour les assurés nés à partir de 1953. Le principe de cette mesure est de permettre un calcul et un versement unique pour les *poly-pensionnés,* c'est-à-dire les

17. Liquidation unique des régimes alignés, dispositif institué par l'article 43 de la loi sur les retraites de 2014.

personnes ayant été affiliées au cours de leur carrière à plusieurs des trois régimes dits *alignés*[18]. Le RSI (aujourd'hui SSI) et la MSA ayant été alignés sur le régime général à partir de 1973, la proximité des règles de calcul de la pension entre les trois régimes a facilité leur intégration dans le modèle en termes de programmation. En revanche, l'implémentation des régimes *non-alignés* a été plus complexe.

La seconde étape fait suite au **projet de régime unique de retraite en points** (et non plus en annuités comme actuellement au sein de l'essentiel des régimes de base) lancé en 2018. Cela impliquait pour Prisme de renforcer sa capacité à simuler des retraites tous régimes, en modélisant plus finement tous les régimes de base ainsi que les retraites complémentaires.

> *« Prisme était désormais en mesure de simuler l'ensemble du régime de retraite en France. »*

L'ampleur de la tâche était cette fois-ci toute autre, dans la mesure où il devenait nécessaire de comprendre, assimiler et retranscrire les réglementations subtiles propres à chaque régime : fonctionnaires territoriaux, hospitaliers ou d'État, régimes spéciaux (SNCF, RATP, industries électriques et gazières, etc.), professions libérales et exploitants agricoles. Une fois apte à déterminer et projeter les trimestres associés à chaque régime, avec les salaires et revenus qui en découlent, ainsi que les dates de départ spécifiques à certains régimes, Prisme était désormais en mesure de simuler l'ensemble du système de retraite en France. Les allées et venues d'un assuré d'un régime à un autre quel qu'il soit ne représentent plus une limite.

Fort de cette faculté, Prisme a alimenté pendant plus d'un an les travaux du Haut-commissariat à la réforme des retraites : en juillet 2019, cette instance a remis au gouvernement son rapport sur la mise en place d'un système universel de retraite (Delevoye, 2019). Un projet de loi portant réforme des retraites a été déposé début 2020. L'étude d'impact menée par la Direction de la Sécurité Sociale pour accompagner le projet de loi est d'ailleurs en grande partie construite sur les résultats issus de Prisme[19].

PENSER LA RETRAITE DE MANIÈRE UNIVERSELLE AU SEIN D'UN SEUL ET MÊME MODÈLE

L'approche du « tous régimes » au sein d'un seul et même modèle a par ailleurs l'avantage de s'exonérer des biais méthodologiques que l'on observe lors de la comparaison entre différents modèles. La mécanique et l'enchaînement des évènements projetés sont pleinement maîtrisés et l'interaction des différentes hypothèses plus lisibles. Les ordres de grandeur ainsi obtenus sont alors plus facilement comparables et les écarts explicables.

Par ailleurs, mieux estimer les pensions versées par l'ensemble du système de retraite permet également d'être plus précis sur le champ initial, le régime général, que ce soit pour calculer l'écrêtement du minimum contributif[20], la condition de ressources pour la réversion ou l'estimation du différentiel avec le plafond de l'ASPA.

18. À savoir la Cnav (régime général), la MSA (Mutuelle sociale agricole) pour les salariés agricoles et SSI (Sécurité sociale des indépendants, auparavant RSI).
19. Voir les références juridiques en fin d'article.
20. Une pension de base au taux plein doit être supérieure à un minimum dit « contributif ». Un complément peut ainsi être versé à un assuré, sous réserve que l'ensemble de ses pensions obligatoires (de base et complémentaires) ne dépasse pas un certain plafond.

Cette évolution de Prisme prend également tout son sens au regard de la multiplicité et de la diversification des demandes en matière d'aide à la décision des politiques publiques. L'expérience de la mise en place de la LURA avec l'intégration des régimes alignés a offert à la direction Statistiques, prospective et recherche de la Cnav une bonne vision des questions que soulève ce type de mesure. Cette expérience construite au fil des années a su révéler une certaine capacité d'anticipation, ou plus modestement de prospective, quant aux interrogations et aux moyens à mettre en œuvre afin de répondre au mieux à ces sollicitations et dans les meilleurs délais.

Au-delà de tous ces aspects techniques, Prisme est aussi une aventure humaine et fédératrice. Cette belle mécanique ne se conçoit pas sans une collaboration de long terme, et elle permet de prendre pleinement conscience de l'apport du travail en équipe et de la pluridisciplinarité. Maintenir tout cela constitue peut-être la principale difficulté dans la pérennité d'une telle structure. Les statisticiens sont maintenant des *data scientists,* la manipulation des bases de données volumineuse tend vers le *big data* et pourtant, derrière ces innovations, il s'agit toujours de la même volonté de fédérer des expertises autour d'un projet commun.

Bien entendu, Prisme entre dans la catégorie des outils encore perfectibles. Pris individuellement, tous les choix et les hypothèses retenus peuvent être débattus. Les méthodologies doivent toujours tendre à plus de pertinence et les améliorations techniques à plus d'efficacité. De même, des décompositions plus fines encore constituent une piste d'amélioration. À l'image de la célèbre cathédrale barcelonaise de Gaudí, Prisme peut d'une certaine manière être perçu comme un éternel chantier.

BIBLIOGRAPHIE

BERTEAU-RAPIN, Caroline, BEURNIER, Paul et DENAYROLLES, Émilie, 2015. La modélisation des trajectoires professionnelles dans le modèle Prisme. In : *Économie et statistique.* [en ligne]. 17 décembre 2015. Insee, N°481-482, pp. 97-120. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/1305195/ES481E.pdf*.

BLANCHET, Didier, 2020. Des modèles de microsimulation dans un institut de statistique – Pourquoi, comment, jusqu'où ? In : *Courrier des statistiques*. [en ligne]. 29 juin 2020. Insee, n°N4, pp. 6-22. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497054/courstat-4-2.pdf*.

BLANCHET, Didier, BUFFETEAU, Sophie, CRENNER, Emmanuelle et LE MINEZ, Sylvie, 2011. Le modèle de microsimulation Destinie 2 : principales caractéristiques et premiers résultats. In : *Économie et Statistique.* [en ligne]. 20 octobre 2011. Insee, n°441-442, pp. 101-121. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/1377517/ES441F.pdf*.

BROSSARD, Cécile, COUHIN, Julie, GRAVE, Nathanël et OLIVEAU, Jean-Baptiste, 2016. Une évaluation des réformes des retraites : quelle sensibilité des résultats aux hypothèses ?. In : *Retraite et société.* [en ligne]. Cnav, 2016/2, n°74, pp. 79-115. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.statistiques-recherches.cnav.fr/images/publications/retraite-societe/Extrait-Brossard-RS74.pdf*.

CCSS, 2020. *Les comptes de la sécurité sociale : résultats 2019, prévisions 2020 et 2021.* [en ligne]. Septembre 2020. Commission des comptes de la Sécurité sociale. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.securite-sociale.fr/la-secu-en-detail/comptes-de-la-securite-sociale/rapports-de-la-commission*.

CHELOUDKO, Pierre et MARTIN, Henri, 2020. Une décennie de modélisation du système de retraite – La genèse du modèle de microsimulation TRAJECTOiRE. In : *Courrier des statistiques.* [en ligne]. 29 juin 2020. Insee, n°N4, pp. 23-41. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497060/courstat-4-3.pdf*.

COR, 2020. *Évolutions et Perspectives des retraites en France.* [en ligne]. 26 novembre 2020. Conseil d'orientation des retraites. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.cor-retraites.fr/sites/default/files/2020-12/Fusion_rapport%2Bsynth%C3%A8se_0.pdf*.

DELEVOYE, Jean-Paul, 2019. *Pour un système universel de retraite.* [en ligne]. Juillet 2019. Haut-commissariat à la réforme des retraites. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://reforme-retraite.gouv.fr/IMG/pdf/retraite_01-09_leger.pdf*.

DI PORTO, Alessandra et GHERNAOUT, Nassima, 2020. La pension de réversion au régime général au fil des générations. In : *Retraite et société.* [en ligne]. Cnav, 2020/1, n° 83, pp. 75-106. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.statistiques-recherches.cnav.fr/images/publications/retraite-societe/RS83-Extrait-Di-Porto_Ghernaout-Pensions-de-reversion.pdf*.

DSPR, 2020. *Présentation du modèle Prisme*. [en ligne]. 5 mars 2020. Cnav, Séance plénière du Conseil d'orientation des retraites, point sur les modèles de microsimulation. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.cor-retraites.fr/sites/default/files/2020-03/Doc%209_%20DSPR_CNAV.pdf*.

FREDON, Simon et SICSIC, Michaël, 2020. Ines : le modèle qui simule l'impact des politiques sociales et fiscales. In : *Courrier des statistiques.* [en ligne]. 29 juin 2020. Insee, n°N4, pp. 42-60. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497070/courstat-4-4.pdf*.

HUMBERT-BOTTIN, Élisabeth, 2018. La déclaration sociale nominative – Nouvelle référence pour les échanges de données sociales des entreprises vers les administrations. In : *Courrier des statistiques*. [en ligne]. 6 décembre 2018. Insee, n° N1, pp. 25-34. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647025/courstat-1-6.pdf*.

LESNARD, Laurent et DE SAINT POL, Thibaut, 2006. Introduction aux méthodes d'appariement optimal (Optimal Matching Analysis). In : *Bulletin de méthodologie sociologique*. [en ligne]. 1er avril 2006. Open Edition, Journals, n°90, pp. 5-25 [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://journals.openedition.org/bms/638*.

FONDEMENTS JURIDIQUES

Décret n° 2012-847 du 2 juillet 2012 relatif à l'âge d'ouverture du droit à pension de vieillesse. In : *site de Légifrance*. [en ligne]. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000026106324*.

Décret n° 2018-390 du 24 mai 2018 relatif à un traitement de données à caractère personnel dénommé « système national de gestion des identifiants ». In : *site de Légifrance*. [en ligne]. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000036940288&categorieLien=id*.

Étude d'impact. Projet de Loi organique relatif au système universel de retraite. Projet de loi instituant un système universel de retraite. [en ligne]. 24 janvier 2020. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.reforme-retraite.gouv.fr/IMG/pdf/etude_d_impact_-_24_janvier_2020.pdf*.

Loi n° 2003-775 du 21 août 2003 portant réforme des retraites. In : *site de Légifrance*. [en ligne]. Modifiée le 24 mai 2019. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/LEGITEXT000005635050/2020-11-28/*.

Loi n° 2010-1330 du 9 novembre 2010 portant réforme des retraites. In : *site de Légifrance*. [en ligne]. Modifiée le 22 janvier 2014. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000023022127/2020-11-28/*.

Loi n° 2014-40 du 20 janvier 2014 garantissant l'avenir et la justice du système de retraites. In : *site de Légifrance*. [en ligne]. Modifiée le 25 décembre 2016. [Consulté le 10 décembre 2020]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000028493476/2020-11-28/*.

QU'EST-CE QU'UNE DONNÉE ?

IMPACT DES DONNÉES EXTERNES SUR LA STATISTIQUE PUBLIQUE

*Pascal Rivière**

Le statisticien public utilise une matière première originale : les données. Mais outre celles qui sont issues d'enquêtes ou de déclarations administratives, il est amené à mobiliser des données d'autres natures, qui ne résultent pas toujours d'un processus d'observation. Comprendre ce matériau « data », c'est en explorer les principales dimensions, en s'appuyant sur le triplet <concept, domaine, valeur>.

Toute donnée se caractérise par un vaste faisceau de conventions (sémantique, nomenclatures, formats, etc.), et par l'infrastructure de connaissances dans laquelle elle s'inscrit, impliquant des choix qui n'ont rien de neutre. Une donnée se révèle aussi dépendante de l'environnement qui lui a donné naissance, et des processus productifs qui l'utilisent. On constate alors que les données ne sont pas pures et parfaites, ne vont pas de soi : paradoxalement, les données ne sont pas données.

Pour les besoins de la statistique publique, utiliser efficacement une telle matière requiert de démêler un entrelacs de conventions, et de construire une sorte d'appareil d'observation a posteriori, rigoureux sur les temporalités, et tenant compte de l'écosystème dans lequel la donnée externe s'inscrit.

Official statisticians use an original raw material, namely data: survey data, but also administrative data. They also use other management data that are not the result of an observation process. Understanding this material means exploring its main dimensions, using the definition of data as a triple <concept, domain, value>.

All data are characterized by a set of conventions, about semantics, classifications, formats, etc. Moreover, data exist within a knowledge infrastructure, and they are stored according to non-neutral choices. Data also depend on the environment in which they were born, and on the productive processes that use them. We then see that data cannot be pure and perfect: data are not given, they are side effects of operational processes.

Using efficiently such a material for the purpose of official statistics requires unravelling the implicit set of existing conventions, and building a kind of observation system a posteriori, taking into account the ecosystem in which these data were embedded.

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

À l'instar de l'ébéniste, du forgeron ou du tailleur de pierre, le statisticien se confronte à un matériau brut, imparfait, traversé de nœuds et de failles. Mobilisant des outils et méthodes qui lui sont propres, il le polit, l'assemble et le met en forme. Ce matériau qu'il travaille, et qu'il contribue à créer, ce sont les données. Or celles-ci pullulent, jaillissent de toutes parts, sans cesse, et dans tous les domaines de la vie : c'est là, semble-t-il, une chance fantastique pour tous les artisans de la donnée. Il s'agit pourtant d'une matière étrange, foisonnante, incroyablement hétérogène : en apparence facilement accessible, elle nous échappe, résistant aux tentatives de définition opérationnelle. Chacun en a sa propre perception, et rétablir la neutralité de l'observateur n'est pas un vain mot. L'objet du présent article est de fournir des clés pour mieux la comprendre, et de voir en quoi les caractérisations proposées ont un impact sur l'activité du statisticien.

LES DONNÉES EN STATISTIQUE

A priori, le lien entre statistique et données va de soi : *« La statistique est la discipline qui étudie des phénomènes à travers la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation afin de rendre ces données compréhensibles par tous »*[1].

Mais il n'est pas sans intérêt de remonter plus loin dans le temps. Si l'on étudie de plus près la littérature classique sur le métier de statisticien (Volle, 1980), ou de nombreux cours de statistique mathématique, on constate que le mot « donnée », tout en étant indiscutablement présent, est moins central qu'on ne l'imaginerait. Il est souvent question de *résultats d'expériences* : la notion d'expérience, cette fois aléatoire, occupe une place centrale dans une autre discipline, celle des probabilités, dont on connaît les forts liens qu'elle entretient avec la discipline statistique. On évoque aussi fréquemment les *observations*. Dans l'esprit, les données de la statistique sont les résultats d'observations relativement à des variables que l'on a définies, et des traitements ultérieurs qu'on leur a appliqués. Implicitement, on fait appel à tout un appareil d'observation, *via* une expérience scientifique (dans le domaine de l'épidémiologie, par exemple), ou *via* une enquête, entre autres possibilités.

> *« Les données de la statistique sont les résultats d'observations relativement à des variables que l'on a définies, et des traitements ultérieurs qu'on leur a appliqués. »*

Certes, les statisticiens publics s'aventurent depuis des décennies au-delà de ce cadre, en ayant recours à des données administratives (les DADS[2], typiquement). Mais à certains égards, le processus déclaratif présente de nombreux points communs avec le processus d'enquête : en forçant un peu le concept, on peut ainsi considérer que les déclarations administratives relèvent d'une démarche d'observation, même si celle-ci s'effectue à des fins de gestion (Rivière, 2018).

Depuis le tournant des années deux-mille, et plus particulièrement depuis une dizaine d'années, on assiste à un changement de paradigme majeur : il ne s'agit plus uniquement pour le statisticien de construire son processus d'acquisition des données, car il existe dans le monde de multiples sources de données, parfaitement ouvertes (*open data*), ouvertes

1. Cf. Wikipédia, Statistique.
2. Déclarations Annuelles de Données Sociales.

sous condition (données accessibles aux chercheurs), *via* des conventions, ou bien payantes sous diverses formes. C'est donc une matière potentiellement accessible, très riche. Mais on ne sait pas comment elle a été élaborée, et en particulier rien ne garantit qu'elle résulte d'une démarche d'observation.

Or réaliser une enquête, mener une expérience scientifique, organiser un processus déclaratif, donnent une vision très particulière et à vrai dire biaisée de ce que peut être une donnée. Dans chacune de ces situations, il s'agit en quelque sorte de prendre une photo à un instant t, en interrogeant le réel, soit à travers un questionnement, soit en sollicitant le monde physique, avec l'usage de capteurs. L'observation est une façon particulière de recueillir les données qui n'est pas la seule possible. Si les données accessibles résultent d'une autre approche, le statisticien doit le savoir afin d'éviter des erreurs dans la transformation de la matière, dans l'interprétation des résultats.

Tout cela oblige à reconsidérer la manière d'appréhender le mot « donnée », dans toutes ses dimensions.

LA DONNÉE : TENTATIVE DE CARACTÉRISATION

Caractériser le concept de donnée est d'autant plus délicat que nombre d'ouvrages sur le sujet *data* éludent tout simplement la question de la définition. L'étymologie fournit un point de départ original, le verbe *donner* n'étant pas neutre ; en anglais, *datum* et son pluriel *data* sont issus du latin *dare* qui signifie… donner. Pour Howard Becker, ce choix est un accident de l'histoire (Becker, 1952) : on aurait dû pointer non pas *« ce qui a été donné »* au scientifique par la nature, mais plutôt ce qu'il a choisi de prendre, les sélections qu'il a opérées parmi l'ensemble des données potentielles. Pour évoquer le caractère partiel et sélectif inhérent aux données, il eût fallu choisir *captum* plutôt que *datum*.

Essayons naïvement les dictionnaires. On y trouve plusieurs explications assez disparates :

- *« Ce qui est donné, connu, déterminé dans l'énoncé d'un problème »* ;
- *« Élément qui sert de base à un raisonnement, de point de départ pour une recherche » ;*
- *« Résultats d'observations ou d'expériences faites délibérément ou à l'occasion d'autres tâches et soumis aux méthodes statistiques » ;*
- *« Représentation conventionnelle d'une information permettant d'en faire le traitement automatique ».*

Les définitions habituelles naviguent ainsi entre statut, fonction, origine et représentation de la donnée.

Dans la littérature sur les bases de données, on peut citer (Elmasri et Navathe, 2016), ouvrage de référence qui définit brièvement et de façon incidente le concept[3] : *« By data, we mean known facts that can be recorded and that have implicit meaning »*[4]. On met donc en exergue le support, l'importance de la sémantique, et on retrouve l'idée de *fait.*

3. On trouve d'autres caractérisations, qui sont celles de l'informatique, mais on les précisera dans la partie suivante.
4. *« Par données, nous entendons les faits connus qui peuvent être enregistrés et qui ont une signification implicite ».*

(Borgman, 2015) étudie plus en profondeur le sujet et aboutit à la définition suivante : « [...] *data are representations of observations, objects or other entities used as evidence of phenomena for the purposes of research or scholarship* »[5]. On retrouve l'idée de représentation du réel (utilisée pour définir l'information), et on constate avec intérêt qu'on ne se limite pas aux observations. Les mots « entité », « objet » apparaissent, l'entité étant définie ensuite par l'auteur comme « *quelque chose qui a une existence réelle* », matérielle ou digitale. Mais on se place dans le contexte d'une utilisation académique.

> *« Si un fait est faux, il cesse d'être un fait, mais si une donnée est fausse, elle reste une donnée. »*

(Kitchin, 2014) consacre tout un chapitre[6] à explorer le mot « *data* ». Il y explicite par exemple la « matière » dont sont faites les données : elles sont abstraites, discrètes, agrégeables, et ont un sens indépendamment du format, du support, du contexte. Il effectue surtout une distinction essentielle entre *fait* et *donnée* : si un fait est faux, il cesse d'être un fait, mais si une donnée est fausse, elle reste une donnée. Ainsi, les données sont ce qui existe préalablement à l'interprétation qui les transforme en faits, preuves, informations ***(encadré 1)***.

Plus généralement, on peut constituer une pyramide données > information > connaissance > sagesse ***(figure 1)***, où chaque couche précède l'autre, et se déduit de la précédente[7] par un « processus de distillation » (abstraire, organiser, analyser, interpréter, etc.), qui ajoute du sens, de l'organisation, et révèle des liens. Ce que l'on peut imager par la formulation de (Weinberger, 2012) : « *L'information est aux données ce que le vin est à la vigne* »), ou celle de (Escarpit, 1991) : « *Informer, c'est donner forme* ».

VERS UNE DÉFINITION EN TROIS DIMENSIONS...

Une des difficultés posées par la notion de donnée, c'est qu'elle embarque dans le même temps deux sujets bien distincts : d'une part la valeur (nombre, code, chaîne de caractères), que l'on va trouver sur un support quelconque, avec un certain mode de représentation, d'autre part le statut de cette valeur, sa sémantique, ce qu'elle est censée représenter. On mêle ainsi des préoccupations opérationnelles et des considérations plus abstraites.

Car si l'on assimile *donnée* à *valeur,* on bute immédiatement sur une contradiction. Lorsqu'on trouve dans un fichier le nombre 324, de quoi parle-t-on : de la hauteur de la tour Eiffel en mètres ? De la surface d'un champ, en hectares ? De la température de fusion d'un métal, en degrés Celsius ? De la vitesse maximum d'un véhicule, en km/h ? Du nombre d'habitants d'un village ?

Prise isolément, la ligne « 324 ; 3 ; 1889 » n'est qu'une succession de caractères vides de sens, au mieux une suite de trois nombres séparés par le délimiteur point-virgule, mais ce ne sont absolument pas des données. Le simple fait de préciser que ce sont des caractéristiques de la tour Eiffel, à savoir hauteur, nombre d'étages et année d'inauguration,

5. *« Les données sont des représentations d'observations, d'objets ou d'autres entités utilisées comme preuves de phénomènes à des fins de recherche ou d'érudition ».*
6. Voir le chapitre « *Conceptualising data* », pp. 2-26.
7. ... même si l'information ne requiert pas systématiquement de se fonder sur une couche de données (cf. le cri d'un animal alertant de la présence d'un prédateur).

> *« La notion de donnée est ainsi indissociable du concept auquel elle se réfère. »*

change tout et confère à cette série insignifiante de chiffres un tout autre statut : intuitivement, il s'agit bien de *données.* La notion de donnée est ainsi indissociable du concept auquel elle se réfère. Elle dépend également d'autres aspects, par exemple ici l'unité de mesure.

Tout cela nous oriente naturellement vers une définition souvent citée, notamment dans la littérature sur la qualité des données (Olson, 2003 ; Loshin, 2010 ; Berti-Équille, 2012 ; Sadiq, 2013). Dans un des premiers ouvrages majeurs sur le sujet, Redman analyse plusieurs définitions connues, et cherche la plus adaptée, en utilisant pour cela des critères : trois critères linguistiques (clarté, correspondance avec l'usage commun, absence de mention du mot « information ») et trois critères d'usage (applicabilité, possibilité d'introduire une dimension qualité, prise en compte des dimensions conceptuelle et de représentation) (Redman, 1996).

Cette démarche le conduit à une définition bien connue des informaticiens : on définit une donnée comme un triplet (entité, attribut, valeur), l'entité étant une modélisation d'objets du monde réel (physiques ou abstraits), l'attribut étant caractérisé par un ensemble de valeurs possibles, ou domaine. L'importance de ce dernier point va nous amener à reformuler très légèrement la définition de Redman, sans trahir la logique initiale, en proposant de définir une donnée par le triplet suivant[8] :

- le **concept** (par exemple hauteur d'un monument), qui se caractérise lui-même par la combinaison d'un objet (ici, monument) et d'un attribut (hauteur) ;
- le **domaine** des possibles : dans le cas d'un monument, un nombre entier positif, et la spécification de l'unité (mètres) ;
- la **valeur** : pour la tour Eiffel, 324.

Essayons maintenant de tirer le fil de chacune de ces dimensions pour mieux appréhender certaines spécificités de la notion de donnée, en particulier pour un usage statistique.

… LE CONCEPT ASSOCIÉ…

Le concept n'est rien d'autre que la signification supposée de la donnée, ce qu'elle est censée représenter, ce qui peut se matérialiser par une définition. Quelques exemples : surface d'un champ, chiffre d'affaires d'une entreprise, profession d'un salarié, mais aussi taille d'un monument, marque d'un véhicule, cours de bourse d'une action, nombre de buts marqués par un joueur, cotation du risque d'un client, diagnostic d'un patient, etc. À l'Insee, on centralise de telles définitions dans le référentiel RMéS (Bonnans, 2019).

On vérifie, à travers ces exemples, que le concept se définit toujours comme un attribut particulier d'une entité, d'un objet. En statistique publique, dans une démarche traditionnelle d'enquête ou de traitement de déclarations administratives, l'entité en question est souvent un individu (ou un ménage), une entreprise, mais ce peut être aussi un logement, un chantier,

8. La définition est équivalente, on n'ajoute ni n'enlève rien, cela permet simplement de faciliter les développements ultérieurs, en séparant clairement une dimension sémantique (concept) et une dimension plus technique (le domaine de valeurs).

Encadré 1. Quelques clés sur la notion d'information

L'irruption du concept d'information date de 1948, au confluent de plusieurs histoires, avec l'arrivée simultanée du fameux article de Shannon (Shannon, 1948) et de l'ouvrage de Wiener (Wiener, 1948). En introduisant une sorte de grammaire universelle de communication, l'un comme l'autre créent un jeu de concepts et de catégories s'appliquant à des sujets aussi divers que les télécommunications, le contrôle ou le calcul mécanique (Triclot, 2014). La notion d'information émerge ainsi dans un univers de machines, et joue un rôle unificateur essentiel, permettant de jeter des ponts entre des disciplines éloignées, en leur fournissant un vocabulaire commun. Ces travaux novateurs permettent aussi de quantifier l'information : dans la vision de Shannon, il y a en toile de fond une problématique de limites de performance pour la compression des messages et leur transmission, et le recours décisif à une représentation digitale qui permet de créer une véritable théorie du code. Shannon est conscient des limites de la chose, et ne pense pas qu'une seule conception d'information puisse rendre compte de toutes les applications possibles*.

Avec Wiener, la cybernétique fait de l'information une nouvelle dimension du monde physique : elle s'ajoute aux modalités d'explication classiques que sont la matière et l'énergie. Elle fait naître une nouvelle classe de problèmes en physique, en introduisant les processus de traitement de l'information.

Avec la montée en puissance des médias, de l'informatisation, l'usage du mot se banalise, mais sans tendre vers une définition simple ni partagée. On peut néanmoins donner quelques éléments explicatifs utiles.

De manière générale, (Buckland, 1991) identifie trois significations : information-objet (données, documents informatifs), information-processus (l'acte d'informer), et enfin information-connaissance (résultante du processus d'information).

(Floridi, 2010) la caractérise comme un bien ayant trois propriétés :

- non-rivalité : plusieurs personnes peuvent posséder la même information ;
- non-exclusivité : c'est un bien facilement partagé… et restreindre ce partage requiert un effort ;
- coût marginal nul.

Enfin, pour (Boydens, 2020), l'information :

- *« résulte de la construction (mise en forme) d'une représentation (perception du réel) ;*
- *au moyen d'un code ou d'un langage au sens large, à savoir tout système d'expression, verbal ou non* […]*, susceptible de servir de moyen de communication entre objets ou êtres animés et/ou entre machines ;*
- *requiert un substrat physique pour être diffusée, qu'il s'agisse de la vibration de l'air* […]*, d'une feuille de papier* […]*, d'un support électronique* […]*;*
- *doit être interprétée pour être utilisée ».*

** « Le mot « information » a été assigné à différentes significations par différents auteurs dans le champ général de la théorie de l'information. Il est probable que quelques-unes de ces significations se révéleront utiles dans certaines applications pour mériter des études supplémentaires et une reconnaissance permanente. On ne peut guère s'attendre à ce qu'une seule conception d'information rende compte de manière satisfaisante des nombreuses applications possibles de ce champ général »* (Shannon, 1953).

> *« Un objet peut avoir de nombreux attributs, mais parmi ceux-ci, certains jouent un rôle particulier : les traits d'identification. »*

un séjour hospitalier, un lycée. Et dans le libellé même du concept, on a souvent tendance à faire disparaître la référence à l'objet, car ce dernier va de soi.

Ajoutons qu'en toute rigueur, la donnée se réfère non pas à un objet en général, mais à un objet donné, à une instance[9] : ce qui fera sens, ce sera la surface d'un champ *bien précis,* la profession d'une personne *bien identifiée,* le nombre de buts marqués par un joueur *donné,* le chiffre d'affaires de *telle entreprise.*

L'attribut devra lui aussi être précisé, notamment sur le plan temporel : la profession à *telle date,* le nombre de buts marqués *telle saison,* le chiffre d'affaires *telle année,* etc. Un objet peut avoir de nombreux attributs, mais parmi ceux-ci, certains jouent un rôle particulier : les traits d'identification. Ce sont ceux qui permettent d'identifier l'objet sans ambiguïté[10], et donc de distinguer une instance d'une autre : on pense naturellement aux nom, prénom, date et lieu de naissance pour un individu ; à la raison sociale, l'adresse pour un établissement ; pour un séjour hospitalier, ce pourraient être par exemple la date de début de séjour, l'identifiant de l'établissement de santé et l'identifiant de l'individu. Les traits d'identification ont la particularité d'être soit inamovibles (date de naissance), soit rarement modifiés (adresse).

Mais revenons aux objets. Les cas de l'individu ou de l'entreprise présentent des avantages incontestables auxquels on ne pense pas toujours :

- une réelle stabilité dans le temps ;
- des traits d'identification indiscutables, permettant de les repérer, et de les distinguer entre eux, sans ambiguïté ;
- l'existence de référentiels reconnus dans lesquels on trouve ces traits, mais aussi des identifiants reconnus comme références communes (NIR, SIREN[11]), avec des principes d'immatriculation relativement transparents et partagés.

À l'inverse, les données susceptibles d'être obtenues dans d'autres sources peuvent se référer à des objets plus délicats à appréhender, car nécessitant une connaissance métier : la ligne téléphonique, le compte bancaire, le compteur électrique. Elles peuvent même concerner des objets volatils : des données comme le montant d'une transaction, la date d'un accident de circulation, renvoient à des objets (transaction, accident) qui s'apparentent à un événement, et qui n'ont donc pas de consistance temporelle.

Ainsi, pour certains types d'objet, il n'existe pas de population de référence, ayant une certaine stabilité et permettant des comparaisons macroscopiques, des contrôles ou des calages sur marges : par exemple on peut difficilement s'appuyer sur un référentiel d'accidents ou de transactions pour comparer à un total connu, avoir un cadre, une limite.

---

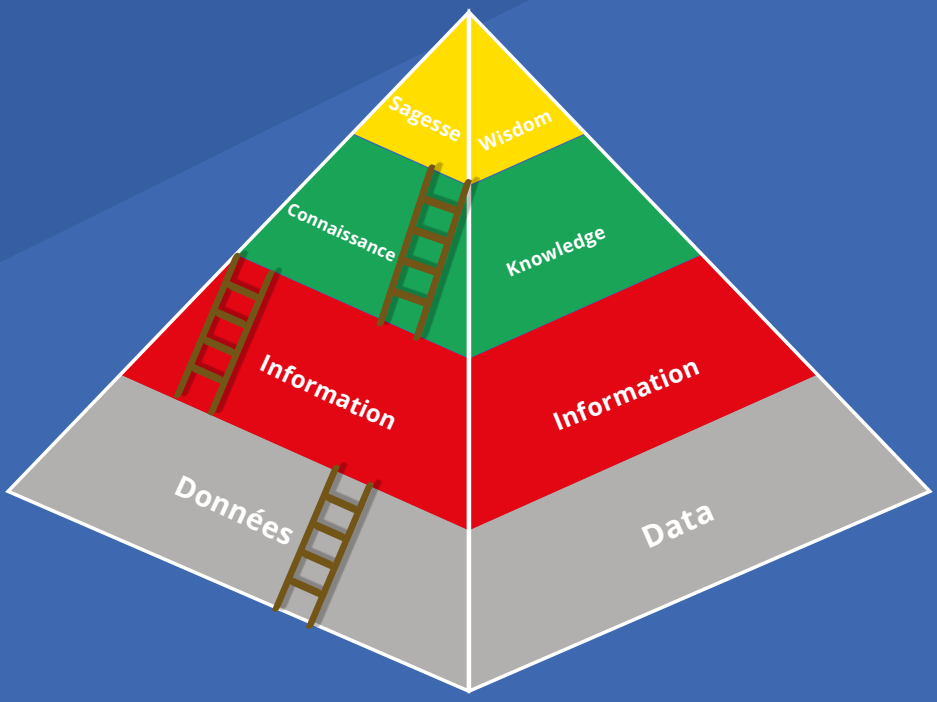
9. Ce terme est fréquent en informatique : on va parler de *classe* (pour l'entité en général) et d'*instance de classe.* Par exemple, Monument est l'entité abstraite, la classe, et la tour Eiffel, ou le Taj Mahal, en sont des instances.
10. Par exemple, (Loshin, 2010) propose plusieurs critères de qualité des données, parmi lesquels figure l'identifiabilité d'un objet (p. 142 et p. 144).
11. Numéro d'identification au répertoire (des personnes physiques) et numéro d'identification des entreprises au répertoire Sirene.

Pour les lignes téléphoniques ou comptes bancaires, les opérateurs ont cette vision mais rien ne les oblige à la partager. De plus, un utilisateur statisticien ou *data scientist* doit pouvoir se ramener à des entités ayant un sens pour l'analyse que l'on veut effectuer : par exemple se ramener au niveau individu, ménage, ou entreprise. Or d'une part il n'y a pas de bijection entre les deux (une personne peut avoir plusieurs comptes bancaires, plusieurs lignes téléphoniques), d'autre part la mise en évidence de ce lien ne va pas de soi (sans parler des problèmes juridiques soulevés).

… LE DOMAINE…

Alors que le *concept* auquel on se réfère reste abstrait, le *domaine* oblige à aborder des considérations plus opérationnelles. (Olson, 2003) le décrit comme l'ensemble des « *valeurs acceptables pour encoder un fait spécifique* ». Il précise que le domaine est générique, indépendant de la manière dont il sera physiquement implémenté. Le définir revient à expliciter les règles que devront respecter les *valeurs*, indépendamment des applications qui vont les utiliser. Si elles sont définies proprement, *via* un référentiel de métadonnées[12], cela facilite considérablement le travail des développeurs des applications utilisatrices[13].

Figure 1. Chaîne de valeur de la connaissance



12. À l'image du référentiel RMéS, décrit dans (Bonnans, 2019).
13. Olson souligne (p. 145) que, dans les faits, peu d'entreprises le mettent en pratique, en prenant l'exemple des dates, des codes postaux et des unités de mesure.

Le domaine dépend de la nature de la valeur, de son type[14]. Si c'est une quantité par exemple, elle sera souvent représentée par un intervalle ; ainsi, on accepte des nombres négatifs pour une température en degré Celsius, mais pas pour la hauteur d'un monument. Elle pourra se référer à une unité de mesure (degré Celsius, mètre), qui fera partie intégrante de la définition du domaine, mais ce n'est pas une obligation, notamment pour les quantités entières (nombre d'enfants).

Dans le cas d'une date, il faudra préciser les formats attendus, de type *jj-mm-aa,* ou *aa-mm-jj,* et on définira souvent l'ensemble des possibles par un intervalle, qui n'a pas tout à fait la même signification qu'un intervalle numérique, étant donné les spécificités des dates (jour ≤ 31, mois ≤ 12).

S'il s'agit d'un code alphanumérique[15], il arrive qu'on définisse le domaine **en intension**[16], par exemple pour les identifiants à structure complexe au sein de grands répertoires (NIR ou numéro de sécurité sociale pour les individus, SIRET pour les établissements d'une entreprise), les numéros de téléphone, et même les noms (Elmasri et Navathe, 2015). Caractériser le domaine par la liste **en extension** aurait en effet deux inconvénients, d'une part la liste serait trop longue à expliciter (il y en a des millions), d'autre part et surtout, elle ne cesse d'évoluer. Et l'on a donc plutôt intérêt à spécifier le domaine sur la base de règles de restrictions de l'ensemble des possibles : par exemple le NIR a une structure très précise, le 1er caractère étant le code sexe, les deux suivants représentant l'année de naissance, puis le mois, etc.[17]

Exprimer l'ensemble en extension est plus fréquent : c'est ce que l'on fait pour un questionnaire en associant codes et réponses (A – Tout à fait d'accord, B – plutôt d'accord, C – plutôt pas d'accord, D – Pas du tout d'accord), ou en discrétisant une variable quantitative (par exemple les tranches d'âge).

> *« Spécifier le domaine revient à le nommer et à définir un certain nombre de propriétés à respecter. »*

Cette liste de codes peut aussi dériver d'une nomenclature… sans en être une. En effet, les nomenclatures comme celles des professions, des activités économiques, des maladies se présentent sous forme d'arborescences à différents niveaux (Amossé, 2020) ; par exemple, dans la nomenclature d'activités française (NAF), les niveaux section, division, groupe, classe, sous-classe, etc. Ainsi, si une donnée a pour concept l'activité économique, le domaine ne pourra être la nomenclature dans son ensemble : ce sera nécessairement un niveau de cette nomenclature, ou plus généralement tout découpage jugé pertinent fondé sur celle-ci, conduisant *in fine* à une liste « à plat ».

On pourrait poursuivre à l'envi avec d'autres types de données, mais dans tous les cas, spécifier le domaine revient à le nommer et à définir un certain nombre de propriétés à respecter : le typage de la donnée, mais aussi une liste de valeurs acceptables, des éléments tels que les règles de restriction de longueur (ou tout autre règle de restriction du champ

14. En toute rigueur, nature et type sont deux propriétés distinctes : par exemple, la CSP est un code à deux chiffres, c'est sa nature. Ainsi, 21 > artisans. On peut représenter ce code par le nombre 21, ou par la chaîne de caractères « 21 » : ce sont deux *data types* différents.
15. Symbole formé d'une succession de caractères qui sont soit des chiffres, soit des lettres.
16. *i.e.* qu'on le définisse par ses propriétés et non par la liste de ses éléments.
17. De telles structures lexicales peuvent être définies mathématiquement par un langage dédié, celui des expressions régulières : voir l'exemple du cahier technique de la DSN (Cnav, 2020), pp. 71-73.

des possibles), les règles relatives à la représentation des valeurs manquantes, voire des points plus techniques comme le jeu de caractères utilisables[18].

Sur le fond, spécifier un domaine revient aussi à effectuer des choix de granularité[19], décisifs pour les usages ultérieurs : la finesse de description dans une nomenclature (cf. l'exemple de la NAF), l'unité de mesure d'une distance (m, km, années-lumière), la précision d'une date (année, jour, minute, seconde, nanoseconde). Cela consiste également à associer aux valeurs admissibles des consignes, des commentaires, indispensables pour créer la donnée (ex. *supra* : si l'on veut dire « plutôt d'accord », on saisira B) et pour l'interpréter (lorsque quelqu'un lira B, il en comprendra le sens).

Avec le *domaine,* on établit ainsi une convention sémantique et technique, commune aux concepteurs et utilisateurs des données, qui est, comme le *concept*, consubstantielle à la donnée. Muni du concept et du domaine, on dispose d'une sorte de réceptacle, au moins théorique, qu'il s'agit maintenant de nourrir d'une *valeur* (ou pas, d'ailleurs).

… LA VALEUR

Avec le troisième élément du triptyque, la *valeur*, on passe à une réalité plus tangible. Il s'agit là du nombre, du code, de la date, de la chaîne de caractères, que l'on associe donc à un domaine et à un concept. Ce dernier est cette fois instancié : la valeur sera relative à une entité et à un attribut bien définis, par exemple la profession et catégorie socio-professionnelle de tel salarié de tel établissement (selon la nomenclature PCS-ESE[20] et au 1er novembre 2020), la masse (en mégatonnes) du soleil, l'altitude (en m) du Mont-Blanc, etc.

Une précision s'impose ici : la valeur est *associée* au concept instancié, elle s'y *réfère*… mais cela ne veut pas dire qu'elle le représente fidèlement, cela n'assure à aucun moment qu'elle soit exacte : on peut trouver des valeurs différentes de l'altitude du Mont-Blanc selon les sources ; un code profession transmis peut être erroné, car il n'est plus à jour, ou parce que la personne qui a renseigné le code s'est trompée ; etc. Ainsi, la question de la fiabilité de la donnée n'est aucunement assurée par le triptyque concept-domaine-valeur. Celui-ci permet de donner une assise à la valeur, de lui conférer un statut autre qu'un simple nombre ou qu'un code vide de sens, mais sa véracité renvoie à d'autres sujets, en particulier à la manière dont la donnée a été construite.

À défaut de certitudes sur la qualité, la référence au domaine peut nous donner une garantie de conformité, une fiabilité de nature syntaxique (Batini et Scannapieco, 2016)[21] : au minimum, la valeur devrait appartenir à celui-ci (par exemple, le fait qu'une donnée représentant une date respecte bien les propriétés que doit avoir une date). Une telle conformité est fréquente… mais non obligatoire : cette fois, tout dépend des contrôles qui auront été effectués automatiquement sur la valeur. Très souvent, les outils de saisie vont être conçus pour que le choix s'effectue parmi des valeurs admissibles présentées à l'écran

18. Voir (Olson, 2003), p. 149, et aussi le cahier technique DSN (Cnav, 2020) évoqués *supra.*
19. On fait aussi un choix de granularité dans le concept, par exemple le fait qu'on se place à un niveau géographique plus ou moins fin.
20. Nomenclatures des professions et catégories socioprofessionnelles des emplois salariés des employeurs privés et publics.
21. Les auteurs distinguent *syntactic accuracy* et *semantic accuracy.* La première est l'adéquation au domaine, indépendamment de la véracité, la seconde est la proximité à la supposée valeur vraie. (Volle, 1980) effectue la même distinction, voir p. 60.

(liste de codes prête), ou bien pour rejeter la valeur si elle n'appartient pas au domaine prévu : c'est ce qu'on trouve par exemple dans la collecte de données par enquêteur. Lorsqu'on a affaire à des échanges de données informatisés et normalisés, le protocole d'échange permet d'assurer des propriétés qui vont même au-delà de l'appartenance à un domaine : c'est le cas des déclarations sociales (Renne, 2018). Plus généralement, les systèmes de gestion de bases de données intègrent implicitement des contrôles de typage. Dans toutes ces situations, par construction, la valeur est donc *conforme,* elle appartient à l'ensemble des valeurs admissibles. Cependant, il peut arriver que les données soient renseignées ou calculées automatiquement sans qu'aucun contrôle ne soit effectué, et soient stockées ensuite dans un fichier, sans garantie de conformité.

Enfin, sur le plan pratique, la valeur se trouve sur un support (physique, logique), qui n'a aucune raison d'être le même que le concept et le domaine. Et elle se réfère à un certain nombre de standards d'encodage qui sont, la plupart du temps parfaitement inconnus des utilisateurs finaux[22].

Comme pour le concept et le domaine, la valeur emporte donc avec elle ses propres règles, qui sont ici de nature plus technique, et des choix techniques auront ainsi été faits : cela vaut pour le type de donnée (*data type*), mais aussi, par exemple, pour la manière de prendre en compte les valeurs manquantes. Ainsi, comme pour les deux autres éléments du triptyque, la valeur rend nécessaire l'existence de conventions. Avec cette troisième et dernière dimension, on constate cependant que raisonner de façon individuelle, donnée par donnée, est largement artificiel.

LA DONNÉE, LES DONNÉES…

> *« La question de la fiabilité de la donnée n'est aucunement assurée par le triptyque concept-domaine-valeur. »*

En effet, le matériau « donnée » a ceci de particulier qu'il ne peut être envisagé seul, tel un grain de donnée séparé du reste. (Borgman, 2015) explique que « […] *data have no value or meaning in isolation; they exist within a knowledge infrastructure – an ecology of people, practices, technologies, institutions, material objects, and relationships* »[23]. Pour leur attribuer un sens, pour en extraire de l'information, il faut les mettre en regard de leurs congénères proches afin d'effectuer des comparaisons, de disposer de l'environnement nécessaire à l'interprétation.

Par ailleurs les données se présentent toujours de façon groupée, elles fonctionnent en meute, en quelque sorte. La plupart du temps, il s'agit de plusieurs instances du même concept, pour un même attribut (par exemple n individus, et pour chaque individu, le code profession), ou de plusieurs attributs de la même instance d'objet (pour un certain individu, toutes les données de déclaration fiscale), ou de plusieurs spécifications temporelles de la même instance, pour un même attribut (par exemple, l'effectif d'une certaine entreprise, année par année).

22. Pour les caractères, les standards ASCII et EBCDIC, par exemple.
23. *« Les données n'ont ni valeur ni sens isolément, elles existent à travers une infrastructure de connaissances – une écologie de personnes, pratiques, technologies, institutions, objets matériels, et relations » .*

> *« Cet enregistrement collectif et non individualisé des données oblige à se donner des règles documentées permettant de retrouver une donnée bien précise au sein d'un vaste ensemble. »*

Cet enregistrement collectif et non individualisé des données oblige à se donner des règles documentées permettant de retrouver une donnée bien précise au sein d'un vaste ensemble de données, et tout ceci dépend de la manière dont les données ont été stockées. Il faut bien les placer quelque part, mais il faut penser en même temps aux moyens de les récupérer.

Il existe pour cela de nombreuses possibilités. Historiquement les données se présentaient dans un fichier structuré, et une technique classique consistait à caractériser chacune d'elles par une plage de colonnes du fichier, en « positionnel fixe »[24], ou par le numéro d'ordre de la donnée dans une liste, en fixe délimité[25]. De telles approches nécessitent toute une documentation associée, en général assez lourde, et complexe à mettre à jour.

On peut procéder autrement, en adoptant un langage de balisage type XML[26], où l'on retrouve en partie l'idée que la valeur doit être enveloppée d'un concept et d'un domaine : chaque valeur figure entre deux balises, ces balises étant elles-mêmes, dans des schémas XML, associées à des règles formelles à vérifier (logique de domaine), et par un nom décrivant le concept. La caractérisation des données est ainsi autoporteuse et ne dépend plus d'une documentation. On peut également citer le format JSON[27], qui est un format plus léger que XML, moins verbeux, efficace, mais en même temps moins riche.

… DANS DES BASES, ENTREPÔTS, LACS, FLUX

Mais la possibilité la plus répandue consiste évidemment à stocker les données dans une *base de données*, gérée par un SGBD (système de gestion de base de données) : ceci induit une forte normalisation, fournit des garanties d'intégrité des données et un langage d'accès aux données (SQL) ; les **bases de données relationnelles** offrent la possibilité d'effectuer de façon fréquente de nombreuses modifications, et s'avèrent donc très adaptées à des processus de gestion (Codd, 1970). Une autre logique, celles des **entrepôts de données**, est en revanche conçue pour faciliter les travaux d'analyse[28], utiles dans le domaine de l'aide à la décision : ce sont alors des données figées, qui requièrent de respecter un cadre contraignant, avec des axes d'analyse communs et donc en particulier des nomenclatures communes alors que les sources d'information sont multiples. La technique des **lacs de données**, là aussi sur données figées, est bien moins contraignante en conception que les entrepôts de données, mais tout le travail de normalisation doit s'effectuer au moment de l'accès aux données.

24. Positionnel fixe : la position d'une donnée est indiquée par les colonnes de début et de fin entre lesquelles elle se trouve, par exemple un prénom entre les colonnes 31 à 50.
25. Fixe délimité : les valeurs sont séparées par des délimiteurs, et on saura par exemple que la donnée que l'on cherche est la troisième dans l'ordre.
26. *Extensible Markup Language*, ou « *langage de balisage extensible* » en français.
27. *JavaScript Object Notation.*
28. On parle de *On-Line Analytical Processing* (OLAP), les premiers papiers sur ce sujet datent de 1993 et impliquent de nouveau Codd, le concepteur des bases de données relationnelles.

Il existe même des techniques de structuration des données conçues pour des situations où elles n'ont pas le temps d'être stockées (systèmes temps réel, notamment) : ce sont les systèmes de gestion de **flux de données**, ou DSMS (*Data Stream Management Systems*) (Garofalakis, Gehrke et Rastogi, 2016).

Au total il existe de nombreuses méthodes pour enregistrer des données de façon organisée, les différences de méthodes ne devant pas être perçues comme étant de nature technique, mais plutôt comme fonction de l'usage envisagé des données (gestion, décisionnel, temps réel) et des éventuelles contraintes sur celles-ci (de volume, en particulier). Dans chaque cas on retrouve d'une manière ou d'une autre le concept et le domaine, soit en tant que (méta)donnée, soit dans une documentation liée. Soulignons enfin que le fait de raisonner sur un ensemble de données et non plus sur une donnée fait émerger d'autres notions : tout d'abord, il apparaît une exigence de cohérence d'ensemble, sur le plan structurel et sur le plan sémantique. Le périmètre que représente cet ensemble de données devient un sujet d'intérêt, à analyser en tant que tel et qui concerne pleinement les statisticiens. Enfin, la connaissance de la source d'information associée à cet ensemble de données est un critère de qualité (Loshin, 2010)[29]. Ainsi, on ne peut aborder les données sans questionner l'écosystème dans lequel elles se trouvent.

ENVIRONNEMENT DE LA DONNÉE

Le triplet (concept, domaine, valeur) fournit un cadre utile mais ne suffit pas à épuiser tout ce que transporte avec elle la matière « donnée ».

Les données ne sont pas données, elles n'existent pas dans la nature de façon immanente : leur existence résulte d'un besoin, elles sont imbriquées dans un environnement. Le fait qu'elles soient définies très rigoureusement ne les rend pas pour autant pures et parfaites, car elles sont *intrinsèquement liées à un usage.* Elles ne sont là que pour concourir à un objectif et non pour mettre à disposition du public une information de référence[30].

> *« Les données ne sont pas données, elles n'existent pas dans la nature de façon immanente. »*

(Denis, 2018) cherche ainsi à « *détricoter les fils de la donnée* »[31]. Il présente le cas tout simple d'un décès : ici, concept et domaine sont aisés à définir. Mais il met en évidence les multiples acteurs qui sont concernés par cette information, et les usages variés en fonction de leurs activités : compagnie d'assurances, administration, etc. Par exemple, les assurances ont besoin de justificatifs : le fait qu'une personne soit considérée comme décédée dans leur système d'information signifie qu'il existe une preuve. Mais à l'inverse une personne décédée pourra être enregistrée comme vivante dans ce même système… tout simplement parce que la preuve n'a pas été transmise : dès lors, la donnée ne reflète pas l'information vraie, elle prépare un usage dans le cadre d'un processus de gestion. De manière plus générale, les données souhaitées peuvent être fonctions d'événements dont rien ne garantit qu'ils soient matérialisés, déclarés de façon simultanée (exemple : l'arrêt de l'activité d'une entreprise).

29. Voir cohérence structurelle et sémantique (pp. 137-139) et la source (*lineage*), comme critère de qualité (pp. 135-136).
30. Sauf dans certains cas particuliers … notamment les données statistiques.
31. Voir pp. 18-19.

Autre exemple : lorsqu'une personne part à la retraite, le montant de cette retraite dépend de ses données de carrière. Une bonne partie de celles-ci sont transmises automatiquement[32], mais ce n'est pas la totalité, et il arrive parfois que le futur retraité transmette des éléments de carrière (feuilles de paie notamment) au dernier moment. Donc dans les données de carrière d'un individu qui n'est pas proche de la retraite, il peut rester des « trous » dus au fait que certaines périodes d'activité n'ont pas encore été déclarées : la carrière telle qu'elle figure dans le système de gestion ne correspondra donc pas à la carrière réelle.

Ainsi les données disponibles ne sont-elles pas nécessairement égales à la valorisation du concept, à la « donnée vraie » telle qu'on l'imagine. Les processus de gestion ne cherchent pas à produire « la vérité » ce qui ne veut pas dire que la donnée soit « mauvaise ». Comprendre la donnée, c'est comprendre l'objectif, les règles du jeu des processus qui la font naître.

> *« Les processus de gestion ne cherchent pas à produire « la vérité » ce qui ne veut pas dire que la donnée soit « mauvaise ». »*

Plus généralement, lorsque les données dérivent d'un cadre juridique (ex : sécurité sociale), il existe d'inévitables écarts entre les données telles qu'elles devraient être selon la loi, les données telles qu'elles seraient si l'on avait une observation parfaite du réel… et les données telles qu'elles sont dans les bases de données (Boydens, 2000). Il existe ainsi un écart entre données théoriques et données réellement récupérées, lié à la réactivité par rapport aux événements. Ceci permet d'aborder différemment la « qualité » des données : la qualité, c'est l'aptitude à l'usage[33], et non la justesse.

Il faut donc comprendre d'où elles proviennent, comment elles naissent (capteur, interface de saisie, calcul automatique, etc.) et connaître l'événement déclencheur de la naissance d'une donnée, mais aussi les processus ultérieurs qui s'en servent (exemple : déclenchement d'une prestation, d'un remboursement). On voit alors moins la donnée comme une vérité absolue, mais comme un maillon nécessaire dans une chaîne complexe. Cela peut permettre d'identifier les raisons pour lesquelles une donnée est absente (exemple de la carrière pour un régime de retraite), ou inexacte sans que le processus opérationnel soit déficient (exemple du décès vu par une compagnie d'assurance).

En comparaison des exemples précédents, les modes d'obtention traditionnels des données en statistique, à savoir les enquêtes, apparaissent comme une étrangeté : la donnée n'est pas directement liée à un usage, par un processus de gestion en aval, elle ne fait qu'alimenter le calcul de données agrégées, les *statistiques*, qui constituent justement un but en soi.

LE STATISTICIEN FACE AUX DONNÉES EXTERNES

Dans un monde numérisé où les *data* accompagnent en permanence nos vies et celles de nos organisations, on pourrait avoir tendance à penser que la multiplication des sources de données sur tous sujets rend le travail du statisticien plus facile. Pour reprendre une expression familière, il n'aurait « qu'à se baisser » pour les ramasser.

32. Le tout récent RGCU (Répertoire général de carrières unique) devrait améliorer sensiblement les choses.
33. *« Quality is fitness for use »* (Juran, 1951).

Cette vision est largement erronée, car elle recèle une confusion majeure sur le sens du mot « donnée » : on peut effectivement accéder, plus ou moins aisément d'ailleurs, à des quantités considérables de *valeurs* (nombres, codes, libellés, dates, etc.)… mais pas à l'équivalent en *données*. Car ces valeurs demeurent lettre morte tant qu'on ne dispose pas de clés solides permettant de les interpréter : concept, domaine, mais aussi périmètre suffisamment explicités. On ne peut envisager d'utiliser ces valeurs à des fins d'analyse sur la base de conventions faibles ou inexistantes, de caractérisations vagues, mal assurées, implicites.

Dès lors, la première responsabilité du statisticien face à des jeux de données externes est de révéler et démêler un entrelacs de conventions sous-jacentes à celles-ci (Martin, 2020)[34] : conventions sur les définitions, les objets, temporalités, nomenclatures, formats, valeurs manquantes, etc., tout ce qui va lui permettre de reconstituer le triptyque concept-domaine-valeur, et de caractériser la population couverte. Les conventions existantes sont à rechercher activement, à valoriser, car ce sont des alliées : on pense par exemple aux normes comptables, très utiles aux statisticiens d'entreprise (elles offrent au passage des garanties de qualité), mais aussi aux fondements juridiques des concepts, qui peuvent conduire à asseoir la qualité de la donnée ou au contraire à devoir pallier ses dérives.

Mais ce n'est pas suffisant : il devra passer de données vivantes, évolutives, liées à des processus opérationnels, à des *observations*. Celles-ci devront être figées dans le temps au même instant t (par exemple la situation des entreprises au 31/12/2019) ou sur une même période (par exemple l'activité des entreprises pour l'année civile 2019), et se référer à des objets distincts les uns des autres, *homogènes* entre eux, de même que leurs attributs. Ce travail d'homogénéisation est central dans l'activité statistique, mais à vrai dire il a commencé bien avant la statistique publique, avec le système métrique et l'unification des poids et mesures (Desrosières, 1993).

En d'autres termes, il s'agit de reconstituer *a posteriori* un pseudo-appareil d'observation, à partir de données qui n'ont pas été conçues à cette fin. Pour ce faire, le statisticien public utilise un puissant socle de conventions, largement partagé : répertoires, nomenclatures, définitions de concepts, unités statistiques, conventions de notation, ce qui se traduit par un vaste ensemble de métadonnées. Les nomenclatures font l'objet d'une vaste coordination (celles de l'Insee ou de l'ATIH[35] par exemple), et les conventions relatives aux populations de référence lui permettent de caler, contrôler et comparer les statistiques obtenues.

> *« Il s'agit de reconstituer* a posteriori *un pseudo-appareil d'observation, à partir de données qui n'ont pas été conçues à cette fin. »*

Avec des données externes non maîtrisées, le statisticien va aussi être confronté à une multitude de défauts de celles-ci, qui peuvent le conduire à les remettre en cause, à questionner leur qualité. Et ce, même si elles sont parfaites pour l'usage auquel elles sont assignées. Pour éviter cet écueil classique, il devra comprendre l'environnement opérationnel des données d'origine pour déterminer ce qu'il peut en faire. Cela affectera les étapes de

34. Sur l'importance des conventions, voir pp. 186-191.
35. Agence technique de l'information sur l'hospitalisation.

contrôle automatique et manuel, d'imputation des valeurs douteuses ou manquantes[36]. L'appariement, et donc l'identification d'objets, l'obligera à questionner le sens, la stabilité de ceux-ci, et leurs liens avec les unités statistiques envisagées (Christen, 2012). La validation finale des agrégats et leur interprétation renverra à la signification de la population de référence couverte.

Ces données externes ne sont qu'un intrant du travail du statisticien, qui s'en servira pour élaborer… de nouvelles données, en l'occurrence des données agrégées nommées « statistiques ». En tant que données, celles-ci auront leurs propres caractérisations, avec une exigence de cohérence particulièrement élevée, ce qui met en exergue le rôle essentiel des métadonnées.

EN GUISE DE CONCLUSION

Il n'existe pas de donnée dans la nature. Pas la moindre. Pour l'exprimer en d'autres termes, les données ne sont pas données, il faut les construire, les prendre (*captum vs datum*). Elles requièrent en amont un travail de modélisation, d'abstraction, de spécification des concepts, puis des domaines, avant d'imaginer produire des valeurs. Elles sont dépendantes de choix eux-mêmes liés à des usages. Mais elles existent dans le vaste monde numérique, et il semble logique que la statistique publique s'interroge sur la manière de les utiliser efficacement pour ses propres besoins, quitte à en inventer d'autres qui soient aptes à nourrir le débat public[37].

C'est là un changement profond pour les statisticiens publics, même si ceux-ci ont une longue expérience d'utilisation de données administratives. Jusqu'au XIX^{e} siècle et même au début du XX^{e}, le recensement était la forme majeure de la statistique publique. Dans la deuxième moitié du XX^{e}, c'est l'enquête qui prend une place prépondérante, avec une maîtrise de bout en bout mobilisant tout un arsenal mathématique et technique. Le XXI^{e} siècle, sans renier les autres formes, serait celui des *data* : le statisticien est aussi informaticien, explorateur des contrées numériques, et, de fait, *data scientist*. Mais en relisant (Volle, 1980), on voit que des fondations demeurent : les répertoires, les nomenclatures, les codes, les définitions, les unités statistiques, tout ce qu'on regroupe aujourd'hui sous le vocable de *métadonnées*. L'évolution majeure de l'activité, outre l'importance prise par l'informatique, réside plutôt dans la nécessité de s'immerger dans les métiers d'origine de la donnée pour mieux s'en servir. C'est ce besoin d'ouverture aux processus externes (et pas seulement aux besoins des utilisateurs), cette polyvalence, cette agilité, cette curiosité teintée de grande rigueur, qui vont constituer la marque d'une nouvelle génération de statisticiens.

36. Exemple : dans le cas du statut vital, précédemment cité (Denis, 2018), on aura intérêt à laisser la donnée telle quelle en cas de décès, et peut-être à effectuer une imputation pour une personne présumée vivante au-delà d'un certain âge.
37. Par exemple les statistiques de mobilité pendant la crise sanitaire de 2020, qu'il aurait été extrêmement difficile et coûteux de bâtir avec des enquêtes classiques (cf. l'article de Jean-Luc Tavernier dans ce même numéro).

BIBLIOGRAPHIE

AMOSSÉ, Thomas, 2020. La nomenclature socioprofessionnelle 2020 – Continuité et innovation, pour des usages renforcés. In : *Courrier des statistiques.* [en ligne]. 27 juin 2019. N° N4, pp. 62-80. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497076/courstat-4-5.pdf*.

BATINI, Carlo et SCANNAPIECO, Monica, 2016. *Data and Information Quality – Dimensions, Principles and Techniques.* Springer. ISBN 978-3-319-24104-3.

BECKER, Howard, 1952. Science, Culture, and Society. In : *Philosophy of Science.* Octobre 1952. The Williams & Wilkins Co. Volume 19, n° 4, pp. 273–287.

BERTI-ÉQUILLE, Laure, 2012. *La qualité et la gouvernance des données au service de la performance des entreprises.* Lavoisier-Hermes Science, Cachan. ISBN 978-2-7462-2510-7.

BONNANS, Dominique, 2019. RMéS, le référentiel de métadonnées statistiques de l'Insee. In : *Courrier des statistiques.* [en ligne]. 27 juin 2019. N° N2. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4168396/courstat-2-6.pdf.*

BORGMAN, Christine L., 2015. *Big Data, Little Data, No Data – Scholarship in the Networked World.* The MIT Press. ISBN 978-0-262-02856-1.

BOYDENS, Isabelle, 1999. *Informatique, normes et temps – Évaluer et améliorer la qualité de l'information : les enseignements d'une approche herméneutique appliquée à la base de données «LATG» de l'O.N.S.S.* Éditions E. Bruylant. ISBN 2-8027-1268-3.

BOYDENS, Isabelle, 2020. *Documentologie.* Presses Universitaires de Bruxelles, Syllabus de cours. ISBN 978-2-500009967.

BUCKLAND, Michael K., 1991. Information as Thing. In : *Journal of the American Society for Information Science.* [en ligne]. Juin 1991. 42:5. pp.351-360. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *http://skat.ihmc.us/rid=1KR7VC4CQ-SLX5RG-5T39/BUCKLAND(1991)-informationasthing.pdf*.

CHRISTEN, Peter, 2012. *Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* Springer. ISBN 978-3-642-31164-2.

CNAV, 2020. *Cahier technique de la DSN 2021-1.* [en ligne]. 14 janvier 2020. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *https://www.net-entreprises.fr/media/documentation/dsn-cahier-technique-2021.1.pdf*.

CODD, Edgar Franck, 1970. A Relational Model of Data for Large Shared Data Banks. In : *Communications of the ACM.* [en ligne]. Juin 1970. Volume 13, n° 6, pp. 377-387. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *https://www.seas.upenn.edu/~zives/03f/cis550/codd.pdf*.

DENIS, Jérôme, 2018. *Le travail invisible des données – Éléments pour une sociologie des infrastructures scripturales.* [en ligne]. Août 2018. Presses des Mines, Collection Sciences Sociales. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *https://books.openedition.org/pressesmines/3934?lang=fr*.

DESROSIÈRES, Alain, 1993. *La politique des grands nombres – Histoire de la raison statistique.* Réédité le 19 août 2010. Éditions La Découverte, collection Poche / Sciences humaines et sociales n°99. ISBN 978-2-707-16504-6.

ELMASRI, Ramez et NAVATHE, Shamkant B., 2015. *Fundamentals of Database Systems.* 8 juin 2015. Pearson, 7e édition. ISBN 978-0-13397077-7.

ESCARPIT, Robert, 1991. *L'information et la communication – Théorie générale.* 23 janvier 1991. Hachette Université Communication. ISBN 978-2-010168192.

FLORIDI, Luciano, 2010. *Information – A very short introduction.* Février 2010. Oxford University Press. ISBN 978-0-199551378.

GAROFALAKIS, Minos, GEHRKE, Johannes Gehrke et RASTOGI, Rajeev, 2016. *Data Stream Management – Processing High-Speed Data Streams.* Springer. ISBN 978-3-540-28607-3.

JURAN, Joseph M., 1951. *Quality-control handbook*. McGraw-Hill industrial organization and management series.

KITCHIN, Rob, 2014. *The Data Revolution – Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE Publications. ISBN 978-1-4462-8747-7.

LOSHIN, David, 2010. *The Practitioner's Guide to Data Quality Improvement.* 15 octobre 2010. Morgan Kaufmann. ISBN 978-0-080920344.

MARTIN, Olivier, 2020. *L'empire des chiffres.* 16 septembre 2020. Éditions Armand Colin. ISBN 978-2-20062571-9.

OLSON, Jack E., 2003. *Data Quality – The Accuracy Dimension*. [en ligne]. Janvier 2003. Morgan Kaufmann. ISBN 1-55860-891-5.

REDMAN, Thomas C., 1997. *Data Quality for the Information Age*. Janvier 1997. Artech House Computer Science Library. pp. 227-232. ISBN 978-0-89006-883-0.

RENNE, Catherine, 2018. Bien comprendre la déclaration sociale nominative pour mieux mesurer. In : *Courrier des statistiques.* [en ligne]. 6 décembre 2018. N° N1, pp. 35-44. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647029/courstat-1-7.pdf*.

RIVIÈRE, Pascal, 2018. Utiliser les déclarations administratives à des fins statistiques. In : *Courrier des statistiques.* [en ligne]. 6 décembre 2018. N° N1, pp. 14-24. [Consulté le 1erdécembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647013/courstat-1-5.pdf*.

SADIQ, Shazia, 2013. *Handbook of Data Quality : Research and Practice*. Springer. ISBN 978-3-642-36257-6.

SHANNON, Claude Elwood, 1948. A Mathematical Theory of Communication. In : *The Bell System Technical Journal*. Juillet 1948. Volume 27, N° 3, pp. 379-423.

SHANNON, Claude Elwood, 1953. The lattice theory of information. In : *Transactions of the IRE Professional Group on Information Theory.* Février 1953. Volume 1, n° 1, pp.105-107.

TRICLOT, Mathieu, 2014. *Le moment cybernétique – La constitution de la notion d'information.* Champ Vallon. ISBN 978-2-876736955.

VOLLE, Michel, 1980. *Le métier de statisticien.* [en ligne]. Éditions Hachette Littérature. ISBN 978-2-010045295. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *http://www.volle.com/ouvrages/metier/tabmetier.htm*.

WEINBERGER, David, 2012. *Too Big to Know.* 1er janvier 2012. Éditions Basic books, New York, p.2. EAN 978-0-465021420.

WIENER, Norbert, 1948. *Cybernetics – Or Control and Communication in the Animal and the Machine*. 1961, 2e édition. The MIT Press, Cambridge, Massachusetts. ISBN 978-0-262-73009-9.

Présentation du numéro N5

Le numéro N5 ne pouvait ignorer le caractère spécifique de 2020 : il commence donc par un article du directeur général de l'Insee sur l'adaptation de l'institut, de ses méthodes, au contexte exceptionnel de la crise sanitaire. Le *Courrier* s'intéresse ensuite à des sujets structurants de gouvernance, à travers l'Autorité de la statistique publique, qui tire un bilan de dix années d'existence, et l'expérience récente du Comité du label de la statistique publique. Comment produire des données utiles à la décision publique ? Avec une représentation cartographique d'une grande souplesse, le carroyage permet de mieux appréhender la réalité des territoires. Avec une communication adaptée, les indicateurs de valeur ajoutée des lycées répondent au besoin d'évaluation et de pilotage interne, comme aux attentes des citoyens et des médias. Avec un modèle de microsimulation dynamique sur les retraites, Prisme accompagne le législateur qui veut faire évoluer la réglementation.
Enfin, le dernier article soulève une question simple : qu'est-ce qu'une donnée ? Exploiter ce matériau constitue le cœur de métier du statisticien, mais en mesure-t-il bien toutes les dimensions ?

ISSN 2107-0903
ISBN 978-2-11-151280-1

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE